

Dissertation zur Erlangung des Doktorgrades der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Novel computational methods
for *in vitro* and *in situ*
cryo-electron microscopy



Dmitry Tegunov

aus

St. Petersburg, Russland

2020

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Novel computational methods
for *in vitro* and *in situ*
cryo–electron microscopy

Dmitry Tegunov
aus
St. Petersburg, Russland

2020

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Patrick Cramer betreut, und von Herrn Prof. Dr. Roland Beckmann von der Fakultät für Chemie und Pharmazie vertreten.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Göttingen, den 02.02.2021

.....
Dmitry Tegunov

Dissertation eingereicht am	07.12.2020
1. Gutachter	Prof. Dr. Roland Beckmann
2. Gutachter	Prof. Dr. Patrick Cramer
Mündliche Prüfung am	28.01.2021

Summary

Over the past decade, advances in microscope hardware and image data processing algorithms have made cryo-electron microscopy (cryo-EM) a dominant technique for protein structure determination. Near-atomic resolution can now be obtained for many challenging *in vitro* samples using single-particle analysis (SPA), while sub-tomogram averaging (STA) can obtain sub-nanometer resolution for large protein complexes in a crowded cellular environment. Reaching high resolution requires large amounts of image data. Modern transmission electron microscopes (TEMs) automate the acquisition process and can acquire thousands of micrographs or hundreds of tomographic tilt series over several days without intervention.

In a first step, the data must be pre-processed: Micrographs acquired as movies are corrected for stage and beam-induced motion. For tilt series, additional alignment of all micrographs in 3D is performed using gold- or patch-based fiducials. Parameters of the contrast-transfer function (CTF) are estimated to enable its reversal during SPA refinement. Finally, individual protein particles must be located and extracted from the aligned micrographs. Current pre-processing algorithms, especially those for particle picking, are not robust enough to enable fully unsupervised operation. Thus, pre-processing is started after data collection, and takes several days due to the amount of supervision required. Pre-processing the data in parallel to acquisition with more robust algorithms would save time and allow to discover bad samples and microscope settings early on.

Warp is a new software for cryo-EM data pre-processing. It implements new algorithms for motion correction, CTF estimation, tomogram reconstruction, as well as deep learning-based approaches to particle picking and image denoising. The algorithms are more accurate and robust, enabling unsupervised operation. Warp integrates all pre-processing steps into a pipeline that is executed on-the-fly during data collection. Integrated with SPA tools, the pipeline can produce 2D and 3D classes less than an hour into data collection for favorable samples. Here I describe the implementation of the new algorithms, and evaluate them on various movie and tilt series data sets. I show that

unsupervised pre-processing of a tilted influenza hemagglutinin trimer sample with Warp and refinement in cryoSPARC can improve previously published resolution from 3.9 Å to 3.2 Å.

Warp's algorithms operate in a reference-free manner to improve the image resolution at the pre-processing stage when no high-resolution maps are available for the particles yet. Once 3D maps have been refined, they can be used to go back to the raw data and perform reference-based refinement of sample motion and CTF in movies and tilt series. *M* is a new tool I developed to solve this task in a multi-particle framework. Instead of following the SPA assumption that every particle is single and independent, *M* models all particles in a field of view as parts of a large, physically connected multi-particle system. This allows *M* to optimize hyper-parameters of the system, such as sample motion and deformation, or higher-order aberrations in the CTF. Because *M* models these effects accurately and optimizes all hyper-parameters simultaneously with particle alignments, it can surpass previous reference-based frame and tilt series alignment tools. Here I describe the implementation of *M*, evaluate it on several data sets, and demonstrate that the new algorithms achieve equally high resolution with movie and tilt series data of the same sample. Most strikingly, the combination of *Warp*, RELION and *M* can resolve 70S ribosomes bound to an antibiotic at 3.5 Å inside vitrified *Mycoplasma pneumoniae* cells, marking a major advance in resolution for *in situ* imaging.

Acknowledgements

I am grateful to my wife Rushana for her love and support.

None of this work would have been possible without the steady support from Patrick Cramer. Thank you for believing in me when few did, and for trusting my ideas – I know this was not always easy. The work environment you created has been second to none in resources, creative freedom, and collegiality.

I thank Roland Beckmann for trusting me to complete this work in a city far away, and for his advice and support.

I am thankful to Clemens Plaschka, Julia Mahamid, Radostin Danev, Ben Engel, Youwei Xu, Sandra Schilbach, Christian Dienemann, Goran Kokic, Felix Wagner, Hauke Hillen, Liang Xue, and many others for our fantastic collaborations.

Lab life would not have been the same without coffee and tea breaks with Carrie, Youwei, Simon, Christian, Svetlana, Sandra, Haibo, and Uli.




Publications

G. Kokic*, H. S. Hillen*, **D. Tegunov***, C. Dienemann*, F. Seitz*, J. Schmitzova, L. Farnung, A. Siewert, C. Höbartner, P. Cramer. (2020)

Mechanism of SARS-CoV-2 polymerase inhibition by remdesivir. bioRxiv 2020.10.28.358481.



H. S. Hillen*, G. Kokic*, L. Farnung*, C. Dienemann*, **D. Tegunov***, P. Cramer. (2020)
Structure of replicating SARS-CoV-2 polymerase. Nature, 584, 154–156.

F. J. O'Reilly, L. Xue, A. Graziadei, L. Sinn, S. Lenz, **D. Tegunov**, C. Blötz, N. Singh, W. J. H. Hagen, P. Cramer, J. Stülke, J. Mahamid, J. Rappsilber. (2020)
In-cell architecture of an actively transcribing-translating expressome. Science, 369, 554–557.

D. Tegunov , L. Xue, C. Dienemann, P. Cramer , J. Mahamid . (2020)
Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.7 Å inside cells. bioRxiv 2020.06.05.136341, Nature Methods ('accepted in principle').

F. R. Wagner, C. Dienemann, H. Wang, A. Stützer, **D. Tegunov**, H. Urlaub, P. Cramer. (2020)
Structure of SWI/SNF chromatin remodeller RSC bound to a nucleosome. Nature, 579, 448–451.

W. Wietrzynski*, M. Schaffer*, **D. Tegunov***, S. Albert, A. Kanazawa, J. M. Plitzko, W. Baumeister, B. D. Engel. (2020)
Charting the native architecture of thylakoid membranes with single-molecule precision. eLife 9:e53740.

D. Tegunov  and P. Cramer . (2019)
Real-time cryo-EM data pre-processing with Warp. Nature Methods, 16, 1146–1152.

J. Mahamid, **D. Tegunov**, A. Maiser, J. Arnold, H. Leonhardt, J. M. Plitzko, W. Baumeister. (2019)
Liquid-crystalline phase transitions in lipid droplets are related to cellular states and specific organelle association. PNAS, 116 (34), 16866–16871.

G. Kokic, A. Chernev, **D. Tegunov**, C. Dienemann, H. Urlaub, P. Cramer. (2019)
Structural basis of TFIIH activation for nucleotide excision repair. Nature Communications, 2885.

Table of contents

Erklärung	II
Summary	III
Acknowledgements	V
Publications	VI
Table of contents	VII
1. Introduction.....	1
1.1 Single-particle analysis and sub-tomogram averaging	1
1.1.1 Sources of noise and signal corruption	1
1.1.2 Data pre-processing	4
1.1.3 Reference-based optimization	6
1.2 Artificial neural networks and differentiable programming	8
1.3 <i>Warp</i> and <i>M</i>	9
2. Automated pre-processing of cryo-EM data in <i>Warp</i>	11
2.1 Results	11
2.1.1 Rationale of <i>Warp</i>	11
2.1.2 Overall design.....	11
2.1.3 User interface	12
2.1.4 Motion correction	14
2.1.5 Estimation of local defocus and resolution.....	15
2.1.6 Particle picking with BoxNet	18
2.1.7 Retraining of BoxNet	20
2.1.8 Online pre-processing during data collection	20
2.1.9 Interoperability with other software	21
2.1.10 Pre-processing tomographic data	21
2.1.11 Template matching	22
2.1.12 Software implementation	23
2.1.13 Benchmarking for 2D data	23
2.1.14 Complementarity of <i>Warp</i> with other tools	25
2.1.15 Benchmarking for tilt series data	27

2.2 Methods	28
2.2.1 Spline interpolation on multi-dimensional grids.....	28
2.2.2 Motion model.....	29
2.2.3 Global and local motion correction.....	29
2.2.4 Contrast transfer function estimation in micrographs	30
2.2.5 Estimation of local defocus	31
2.2.6 Resolution estimation	33
2.2.7 Contrast transfer function estimation in tilt series.....	33
2.2.8 Considerations for tomogram reconstruction	35
2.2.9 Tomogram reconstruction	36
2.2.10 Export of corrected data	37
2.2.11 Particle picking with a residual neural network.....	38
2.2.12 Initial training of BoxNet	40
2.2.13 Retraining of BoxNet	42
2.2.14 Template matching in micrographs and tomograms.....	42
2.2.15 Deconvolution	43
2.2.16 Denoising improves particle visibility	44
2.2.17 Benchmarking for 2D data	45
2.2.18 Benchmarking for tilt series data	47
3. Multi-particle refinement in <i>M</i>	48
3.1 Results	48
3.1.1 Overall design.....	48
3.1.2 Multi-particle system modeling	50
3.1.3 Correction of electron-optical aberrations	53
3.1.4 Optimization procedure	56
3.1.5 Map denoising and local resolution.....	58
3.1.6 Contribution of different model parameters to map resolution	59
3.1.7 Similar resolution obtained from frame and tilt series data	61
3.1.8 Comparison with RELION on atomic-resolution frame series data	62
3.1.9 Comparison with other tools for tilt series data refinement.....	63
3.1.10 <i>M</i> enables the visualization of an antibiotic bound to 70S ribosomes at 3.5 Å in cells.....	65

3.2 Methods	69
3.2.1 Data management.....	69
3.2.2 Deformation model.....	70
3.2.3 Imaging model.....	71
3.2.4 Optimization procedure.....	72
3.2.5 Memory footprint considerations.....	76
3.2.6 Avoiding CTF aliasing.....	78
3.2.7 Data-driven weighting.....	80
3.2.8 Map reconstruction.....	81
3.2.9 Map denoising.....	82
3.2.10 Assessment of map denoising.....	84
3.2.11 Acquisition of apoferritin benchmark data.....	85
3.2.12 Comparison between frame and tilt series performance.....	86
3.2.13 Assessment of multi-species refinement.....	87
3.2.14 Comparison with RELION on atomic-resolution frame series data	87
3.2.15 Comparison with other tools for tilt series data refinement.....	88
3.2.16 Acquisition and refinement of <i>M. pneumoniae in situ</i> tilt series data	89
4. Conclusions and outlook.....	92
4.1 Further development of <i>Warp</i> and <i>M</i>	93
4.2 Central repository for sharing <i>in situ</i> cryo-ET data	94
4.3 Resolving compositional and conformational heterogeneity with machine learning.....	95
4.4 Applying machine learning to <i>in situ</i> cryo-ET data.....	96
List of abbreviations	99
List of figures and tables	100
References.....	102

1. Introduction

1.1 Single-particle analysis and sub-tomogram averaging

1.1.1 Sources of noise and signal corruption

Modern cryo-electron microscopy¹ (cryo-EM) can image 2D projections of the 3D Coulomb potential of biological macromolecules ('particles') in vitreous ice with atomic accuracy². However, the precision of such measurements is very low due to several sources of noise³. Some of the sources can be assumed to follow a 0-mean, Gaussian distribution⁴, while others can be explicitly corrected for. Thus, the underlying signal can be retrieved up to a very high resolution by filtering and averaging many independent measurements of parts of the same signal⁵. This is possible because for many species of biological macromolecules, copies of the same molecule exhibit a very high degree of structural similarity. Aligned to a common 3D reference frame, the 2D projections can be related to parts of the underlying 3D potential through the central-slice theorem. This enables the averaging of projections from arbitrary directions to obtain a 3D reconstruction of the Coulomb potential with significantly increased signal-to-noise ratio (SNR)⁵.

Biological macromolecules are fragile and cannot sustain a large amount of electron radiation in the microscope⁶. While any sampling the molecule with electrons destroys its structure by knocking individual atoms from their original positions in random directions, it is a gradual process. The effect can be averaged out given enough data, but explicit exposure-weighting can improve the data efficiency⁷. Typical micrographs use an overall exposure of 40-50 e⁻/Å², and fractions of this exposure are filtered *in silico* in the frequency domain to down-weight the high-resolution part of the signal as a function of the accumulated exposure⁷. In the weighted data, only the first 2-4 e⁻/Å² contribute meaningfully to the signal at 2 Å resolution and beyond. The sampled object's phase contrast as well as parameters of the electron-optical system determine the 2D probability function of where the electrons will impact the detector. For every electron, the time and position of its impact on the direct electron detector (DED) sensor are recorded. DEDs

significantly improve the event localization and secondary-scattering noise compared to previously used indirect CCD-based solutions⁸, now approaching the theoretical maximum detective quantum efficiency (DQE). However, even with a $40 \text{ e}^-/\text{\AA}^2$ exposure on a DED, the measured probability function remains vastly undersampled, leading to the most significant source of noise: shot noise³. Although it follows a Poisson distribution, its long tail is ignored and a Gaussian distribution is assumed in all current processing methods.

Large-scale sample motion occurs during the exposure as well^{9, 10}. Mechanical stage instabilities lead to fast, global shifts of the entire field of view throughout the exposure. In a tilt series, the assumed stage angles are also inaccurate because of stage instability. Local shifts occur at a slower pace due to beam-induced motion (BIM)⁹. The fast readout speed of a DED allows to fractionate the exposure and thus the motion into frames, and align them later *in silico*¹⁰. If the motion is fast or the fractionation is coarse, some intra-frame motion remains. This decreases high-resolution signal anisotropically and dampens its amplitude in the final average if left uncorrected.

Because biological macromolecules consist of elements of similar mass as the surrounding vitreous ice, amplitude contrast is negligible in cryo-EM. Instead, phase contrast¹¹ between elastically scattered and unscattered electrons is achieved through deliberate introduction of aberrations (typically defocus), or a phase plate¹² in the electron-optical system. Because the aberration-induced phase shift grows with increasing spatial frequency, the resulting contrast transfer function¹¹ (CTF) oscillates between -1 and 1 in the frequency domain. With all other parameters fixed, the defocus determines how fast the CTF oscillates and where the contrast peaks and reversals are in the frequency domain. In the spatial domain, this determines how far sidebands of a signal corresponding to a certain spatial frequency are offset from its original position¹³. Further higher-order patterns in the CTF's phase shift come from astigmatism, trefoil and other electron-optical aberrations, and can be modeled with Zernike polynomials¹⁴. Because the specimen rarely lies flat in the focal plane, defocus variations can be expected within a micrograph

and even between individual macromolecules. CTF corrupts the image signal in a way that cannot be averaged out because signal randomly modulated with -1 and 1 will average to 0. Thus, all signal observations must be demodulated using the respective CTF before they can be averaged. Many of the CTF parameters are not known *a priori* and must be determined by fitting an analytical model to the signal. The CTF modulates the particle signal and solvent noise, but not the shot noise³.

The particles of interest are embedded in a thin layer of vitreous ice. Similarly to the particles, its water molecules scatter electrons and are present in the images. Because all but a few water molecules around each particle are located randomly, their contribution to the projections can be treated as random Gaussian ‘solvent noise’³ that will average out. Thicker ice results in stronger solvent noise, but also leads to more inelastically scattered electrons, which are an additional source of noise. Impurities of other compounds are also located within the ice layer and on its surface¹⁵. Some of them, such as ethane drops, are visible as high-contrast spots in the image. Such areas are best excluded from averaging because they break the Gaussian noise assumption. Macromolecules can also overlap with each other’s projections in overly concentrated *in vitro* samples, or when imaging the crowded cellular environment¹⁶. From the point of view of averaging, overlapping projections are additional solvent noise that is assumed to have a Gaussian distribution in all current algorithms.

A very high degree of structural similarity between copies of the same molecule at the time of vitrification is essential for averaging. If extensive classification and alignment¹⁷ of conformationally or compositionally heterogeneous regions cannot reduce the heterogeneity, the resulting average will be a superposition of all present states, limiting the resolution. For compositional heterogeneity (‘occupancy’), the resulting noise follows a bimodal distribution⁴. Although the superposition can reach high resolution, it might be hard to disentangle without knowing the underlying atomic model. Conformational heterogeneity (‘flexibility’) often results in approximately Gaussian noise and anisotropic

amplitude attenuation. Given vast amounts of data, high resolution can be obtained and the amplitude attenuation reversed. However, collecting so much data is impractical.

In summary, single-particle analysis⁵ (SPA) and the closely related sub-tomogram averaging¹⁸ (STA) techniques aim to model the noise in the data accurately, and to correct it as far as possible. With only 0-mean, Gaussian noise left in the data before averaging, modern hardware and processing pipelines can reach 3 Å and better with 10^5 - 10^6 asymmetric particles^{19, 20}.

1.1.2 Data pre-processing

Before they can be fed into SPA/STA tools, cryo-EM data need to be pre-processed. Although the necessary algorithms are accessible through SPA/STA software suites, they are not strictly part of the SPA/STA concept and often available as stand-alone tools.

Cryo-EM samples experience global and local motion during exposure. If the exposure is fractionated into many parts, this motion can be largely canceled out. Reference-free alignment algorithms²¹⁻²⁵ attempt to estimate a shift vector for each frame in the exposure that minimizes the difference between each aligned frame and their average. To account for local motion, the field of view can be split into quadrants, or a field of several vectors can be estimated per frame with smooth interpolation between them. Because each frame and the resulting average have very low SNR, the fitted model requires strong regularization. This puts a limit on the temporal and spatial resolution achievable for the motion model with reference-free alignment. In tilt series, the different projection angles in different parts of the exposure pose an additional problem because the overlap in underlying signal is very small, making alignment more difficult. In the absence of a priori knowledge of particles, high-contrast features such as gold fiducials or patch-tracking algorithms are used. Because of the increased difficulty, attempts at local motion alignment usually bring little benefit²⁶.

Without the ability to perform CTF correction, averages of cryo-EM data would get stuck at ca. 20 Å, i.e. where the first contrast reversal occurs on the average. Thus, any kind of

further analysis requires a priori knowledge of the CTF parameters for each image. Most of the parameters, such as acceleration voltage, spherical aberration, or pixel size, can be treated as fixed at this point. Defocus, astigmatism and, in the presence of a phase plate, phase shift are different in every image and must be fitted. This can be done in a reference-free manner by fitting a simulated CTF to the amplitude or power spectrum of the image that is modulated by the CTF²⁷⁻²⁹. Because the spectrum does not capture the phase and is strictly positive, the CTF's absolute value or its square is used. The contribution of shot noise and inelastic scattering to the spectrum is significant and varies between spatial frequencies. This can affect the fitting, and even break it completely. Thus, prior to fitting the CTF, the spectrum envelope must be fitted. The lower boundary should ideally go through all zero-crossings of the CTF, while the upper boundary should go through its peaks. The lower boundary is then subtracted from the spectrum to improve the fitting²⁷. In some cases such as very low defocus, the envelope estimation itself can break. Thus, CTF fitting algorithms are not entirely robust, and may require some supervision. It is desirable to fit the defocus more locally to account for an uneven or tilted sample³⁰, but reference-free per-particle estimation does not have enough signal to deliver accurate results.

One of the central assumptions in SPA, and STA by extension, is that each piece of data contains a single particle roughly at its center³¹. To satisfy this condition, particles must be located and extracted from motion-corrected micrographs or tilt series. Due to low SNR and the presence of other molecules and high-contrast artifacts in the images, this is often a difficult task. In addition, bias in the selection procedure will affect all downstream analysis. Template-free methods^{32,33} attempt to find features within the selected size range, while excluding bigger and smaller features. Although the methods do not introduce bias beyond the object size, they can have very low specificity in the presence of significant noise and other objects in the images. Template-based methods³⁴⁻³⁶ require at least a set of low-resolution templates, which can be obtained by processing the results of template-free methods or manual selection. Using these positive examples, the

methods are more accurate, but still pick up high-contrast artifact because the algorithms lack knowledge of the negative example space (i.e. ‘what is not a particle’). If the particle selection is unreliable, extensive manual curation is required, which can take several days for a few thousand micrographs.

1.1.3 Reference-based optimization

Once located and extracted, individual particles must be iteratively aligned to a 3D reference so they can be averaged to obtain a high-SNR reconstruction⁵. In case of a heterogeneous sample, each particle must be compared to multiple references (‘classes’) to find the one it most likely belongs to. During alignment, the most likely pose, i.e. orientation and in-plane translation, is estimated³⁷. The reference is projected (in 3D in case of STA³⁸) at the sampled angles and offsets, and an imaging model is applied to these forward-projections. The model includes modulation with the estimated CTF, and a weighting function based on the estimated spectral SNR (SSNR) to give less weight to the noisier parts of the signal³⁷. The projections are then compared to the experimental particle image to select a projection that matches it the most, and the corresponding pose is assumed to be optimal in the current iteration. Once poses have been established, particles are CTF-corrected, weighted, and back-projected in 3D. The result is divided by the back-projected sum of weights to obtain the weighted average as the reconstruction³⁷. If the resolution of this reconstruction is better than the previous reference, it can be used for an additional reference-based alignment iteration with the prospect of improving the resolution further. The resolution is estimated through Fourier shell correlation (FSC)³⁹, which is a form of cross-validation and requires the particle set to be split in 2 halves that are refined independently.

Even the first iteration of reference-based alignment requires a reference. In some cases, the structure of a similar macromolecule is available, which is enough to guide the optimization in the right direction. If no initial structure is available, the optimization starts with one or several random low-resolution blobs. Stochastic gradient descent (SGD) is a popular method⁴⁰ used to slow the convergence rate of the optimization and decrease

the chance that it will get stuck in a local optimum. The algorithm aligns a small random subset of the particles to the current reference in each iteration, and updates the reference towards being more similar to the resulting reconstruction. Coupled with SGD, STA can improve the convergence even further because each sub-tomogram covers a larger part of the underlying signal and contains information about the correct handedness of the structure, which is degenerate in SPA data⁴¹.

By going back to the raw micrograph data, reference-based optimization can improve the previous reference-free frame alignments⁴². Instead of optimizing the similarity between each frame and their average, forward-projections of a high-resolution reference are used. The projections can be either combined to a larger image and compared against entire frames⁴³, or compared to particle images extracted from frames⁴². Because the projections contain far more signal than frame averages, the granularity of the motion model can also be increased to fit faster and more localized movements. The same approach can be applied to tilt series that were previously aligned through fiducial or patch tracking⁴³. Instead of comparing 3D forward-projections against sub-tomograms, a set of 2D projections is compared against particle images extracted from all tilt images at positions corresponding to the sub-tomograms position in 3D. This way shifts and angles can be optimized for individual tilts.

The CTF is another part of the image model that can be improved with reference-based optimization. Instead of sampling different particle poses, values such as defocus and astigmatism are changed in the CTF the forward-projections are modulated with to find the combination with the highest similarity to the particle images. As with frame alignment, this provides better signal than fitting against the amplitude spectrum¹⁴. The fit is also not affected by a contrast-rich substrate such as carbon, which may lie above or below the particles and systematically bias the defocus. Reference-based optimization enables the fitting of defocus and, in some cases, astigmatism per-particle¹⁴. Symmetric and asymmetric higher-order optical aberrations can also be estimated based on reference projections by altering the corresponding Zernike polynomial factors in the CTF.

1.2 Artificial neural networks and differentiable programming

Inspired by a simplified model of biological neurons, artificial neurons accumulate input signals and apply a non-linear sigmoid function to the result⁴⁴. Composed of multiple layers of neurons, neural networks can approximate any mathematical function. To achieve this, the weights each neuron assigns to its inputs must be adjusted⁴⁴. For most practical uses, the true function that needs to be approximated is not known, and only pairs of input and output values are available. Because any non-trivial function requires thousands or millions of neurons for accurate approximation, an exhaustive parameter search is not feasible. Instead, parameters are ‘trained’ in a gradient descent-like procedure⁴⁵ using gradient information obtained by presenting the network with an input, and comparing its output to the target output.

Modern neural networks⁴⁶ are far more versatile than the original ‘perceptron’ concept⁴⁴. Instead of only accumulating inputs and applying sigmoid activation, any number of differentiable operations can be concatenated in intricate graphs. The differentiability is crucial because every output variable can then be related to every input variable through a fully differentiable path, and gradients for all parameters involved can be computed easily. Hence, this approach to building and training machine learning models is called ‘differentiable programming’⁴⁷. The gradient is calculated in a process called ‘back-propagation’⁴⁸, which uses the chain rule of differentiation to propagate the gradient back from the output layer, operator by operator. Without this algorithm, training large networks would not be computationally feasible because the network would need to be evaluated at least once for each parameter to obtain the gradients.

Convolutional neural networks (CNN)⁴⁹ are particularly popular for computer vision tasks⁴⁶. In the ‘fully connected’ layers of classical neural networks each neuron is connected to every output variable of the previous layer. In CNNs, a small kernel of weights (e.g. 3x3 pixels in case of image data) is applied at each position in the input data and the results are summed, convolving overlapping groups of spatially adjacent variables. A non-linear function is applied to the result of each convolution, and they are combined in a

layer of the same dimensionality as the input. Multiple kernels are usually applied to the same input to obtain multiple channels in the output layer. Kernels in the next layer then extend in the additional channel dimension. The initial convolutions act as edge detectors, while subsequent convolutions can capture larger and more complex features. Unlike fully connected networks that learn to expect an object at a certain position in the input data, CNNs are translation-invariant and will produce the same encoding for an object at any location. Over the past decade, CNNs have outperformed every other approach to segmentation⁵⁰, image classification⁵¹, denoising⁵², super-resolution⁵³, and many other tasks.

1.3 Warp and M

In this thesis, I describe and evaluate two new software tools developed to tackle two major problems in cryo-EM data processing: robust and automated data pre-processing, and comprehensive reference-based optimization of the imaging model in movie and tilt series data. Used in conjunction with an established SPA tool like RELION, the new algorithms provide a highly automated pipeline that achieves record-high resolution for challenging samples.

Warp takes over all pre-processing steps such as motion correction, CTF estimation, particle picking, and tomogram reconstruction. The new tool integrates its algorithms into a fully graphical, intuitive user interface that helps the microscopist with immediate feedback during data collection. The new algorithms can reliably handle modern data acquisition regimes such as tilted movie collection, and dose-symmetric tilt series. The application of CNNs in conjunction with an extensive training data corpus enables very robust particle picking, while micrograph denoising greatly helps the user in making sense of low-contrast images. Operating in parallel with automated data acquisition, *Warp* provides a constantly updated stream of accurately picked particles that can be immediately fed into 2D classification or *ab initio* 3D refinement to quickly assess the sample quality and, in favorable cases, complete most of the processing by the time data acquisition is finished. To evaluate its performance, *Warp* is used to pre-process several challenging

data sets. Most notably, a previously published data set that produced a 3.9 Å influenza hemagglutinin trimer map is pre-processed with *Warp* and cryoSPARC in a highly automated fashion to obtain an improved 3.2 Å map.

M treats the contents of a micrograph or tomogram as a physically connected multi-particle system instead of following the single-particle assumption from SPA. The new multi-particle framework enables the simultaneous reference-based optimization of all aspects of the sample and imaging model such as particle poses, physically plausible sample and particle motion trajectories, and CTF parameters including higher-order aberrations. The optimization procedure can incorporate multiple references of different resolution at the same time, thus benefitting from the signal of more particles per field of view in heterogeneous samples. *M* unifies the processing of movies (further denoted as ‘frame series’) and tilt series by treating the latter as series of images with angular and translational constraints imposed by the multi-particle sample model. Evaluation shows that the new approach can achieve the same high resolution with frame and tilt series data of the same sample. *M* further integrates machine learning-based map denoising to robustly filter its reconstructions to local resolution and to avoid the overfitting of low-resolution regions. The new tool is evaluated on several published frame and tilt series data sets, and improves the resolution in each case. Most strikingly, *M* is tested on a tilt series of *Mycoplasma Pneumoniae* cells and is able to resolve the 70S ribosome bound to an antibiotic at 3.5 Å, paving the way for high-resolution structural biology inside cells.

2. Automated pre-processing of cryo-EM data in *Warp*

The work presented in this chapter was published in:

D. Tegunov, P. Cramer. Real-time cryo-electron microscopy data preprocessing with *Warp*. *Nat Methods* **16**, 1146-1152 (2019).

2.1 Results

2.1.1 Rationale of *Warp*

We aimed at providing a software package that enables the electron microscopy user to evaluate, correct, and process cryo-EM raw data immediately during data acquisition. The rationale was to provide a single, streaming interface between the data acquisition software that produces the raw data, and the existing software solutions for 2D classification and 3D refinement of pre-processed cryo-EM single-particle data, such as RELION³⁷ or cryoSPARC⁴⁰. We called the resulting software package '*Warp*' in reference to its fast correction of object distortions that occur in cryo-EM, and its GPU-based implementation that results in almost instantaneous output. Our rationale was that *Warp* should be used for the online evaluation, correction and processing of both 2D and tilt series cryo-EM data. *Warp* can be installed on standard platforms and operated by non-expert users via a streamlined user interface (UI) that has been developed in parallel to the underlying algorithms to augment their operation. *Warp* was designed to be widely applicable for biological data acquisition at any cryo-EM facility and substantially speeds up the process of cryo-EM structure determination with improved results.

2.1.2 Overall design

A schematic of the computational steps carried out by *Warp* is provided in **Figure 2.1**. For simplicity, we first describe the workflow for 2D data, before we describe the application to tomographic tilt series at the end of the results section. At the beginning of the pipeline, *Warp* reads any new data saved by the acquisition software. *Warp* then estimates and corrects the motion captured in the frames both globally and locally. Next, *Warp* fits

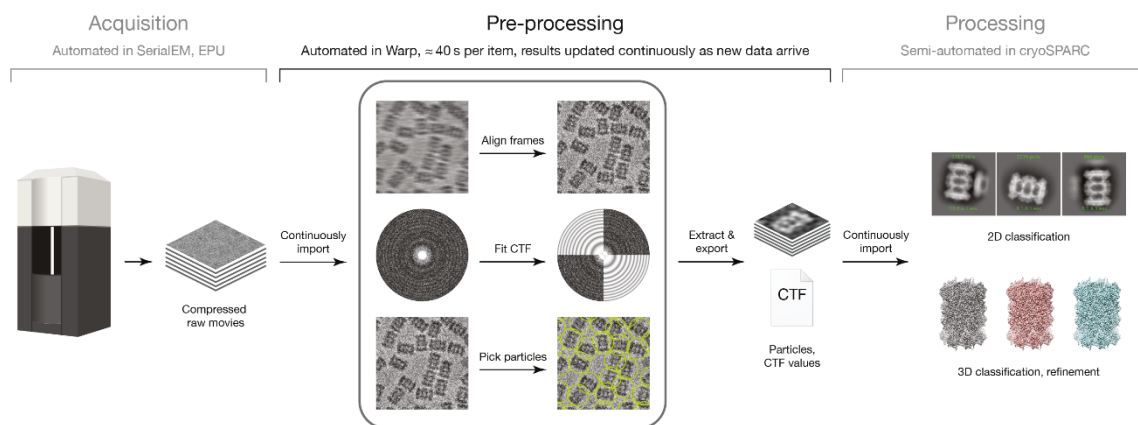


Figure 2.1 | Warp handles all pre-processing steps to close a gap in the 2D cryo-EM pipeline.

Data is acquired by the microscope in an automated fashion and stored as compressed stacks of movie frames. Warp continuously monitors its input folder for new files, and subjects them to all steps of the pre-processing pipeline: frame alignment, CTF estimation and particle picking. Warp writes out a stack of particles for each pre-processed micrograph and maintains a dynamically updated STAR file with references to all particles and their local CTF parameters. This file can be used as a data source in a tool such as cryoSPARC, which will periodically run subsequent processing steps like 2D classification and ab initio reconstruction on the latest set of particles.

a spatially resolved CTF model, enabling the assignment of local defocus values to any particles extracted from the micrograph later. *Warp* then uses a neural network-based approach to automatically pick particles from the corrected micrographs with very high accuracy. Finally, *Warp* exports the resulting dose-weighted particle images to a downstream structure determination program such as RELION³⁷ or cryoSPARC⁴⁰, which carry out 2D and 3D classification, map refinement and reconstruction. During pre-processing, *Warp* provides a comprehensive overview of all important data parameters, allowing the operator to tune the acquisition settings to achieve optimal results faster. In the following we will describe the most important components in more detail.

2.1.3 User interface

Warp's UI is designed to help the user to comprehend and interact with the thousands of data objects generated routinely during cryo-EM data collection (**Figure 2.2**). The 'Overview' tab displays various properties, such as defocus, estimated resolution, amount of

DURCHSCHNITTSELEKTRONNENMIKROSKOPIEBILD-DATENENTZERRUNGSWERKZEUG 1.0.6

SAVE SETTINGS LOAD SETTINGS

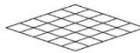
Input
 Input: [Z:\data\dingund\tempier_10097\Raw-frames\31800... -- ".mic](#)
 Pixel X/Y: 1.3100/1.3100 Å \leftrightarrow 0.0 °
 Binc: 0.00x (1.3100 Å/px)
 Dose: 0.82 e/Å²/frame

Preprocessing
☒ Correct gain using: [Z:\data\dingund\tempier_10097\Raw-frames\31800... -- ".mic](#)
☐ Flip X axis ☒ Flip Y axis ☐ Transpose


☒ CTF
 Window: 768 px Range: 0.05–0.44 Ny ☐ Use Movie Sum
 Voltage: 300 kV C_d: 3.70 mm C_s: 2.70 mm
 Amplitude: 0.07 IL Aperture: 30 µrad ΔE: 0.70 eV
 Defocus: 0.2–8.0 µm ☐ Phase Shift

☒ Motion
 Consider 0.03–0.60 Ny, weight with B = -16.0 Å²


Models



Defocus: 0 x 6 x 1



Motion: 5 x 5 x 60


☒ Pick Particles
 Use [BoxNet2Mask_20180918](#)
 Expect 130 Å [vpy](#) particles; use scores above 0.30
 Maintain a minimum distance of 0 Å from 
☐ Extract 480 px boxes, 1.3100 Å/px ☒ invert ☒ normalize
☒ Maintain a separate list of the latest 10000 particles

Output
 Skip first 2, last 0 frames
☒ Average
☐ Aligned stack, collapse every 1 frames

Overview Fourier Space Real Space


EXPORT MICROGRAPH LIST ADJUST PARTICLE DEFOCUS EXPORT PARTICLES IMPORT PARTICLE COORDINATES MATCH TEMPLATE EXPORT BORNET EXAMPLES


Processing Status



124
317

Astigmatism (use up to 3.0 o)




Defocus (use 0.35–3.30 µm) — average |CTF|: 

Estimated resolution (use better than 6.0 Å)

Average motion per frame in first 1/5 (use up to 3.0 Å)

Number of particles in [_BoxNet2Mask_20180918](#) — 435446 overall, 321924 good (use at least 200)

 (use up to 10 %)

STOP PROCESSING

Overview Fourier & Real Space

PROCESS ONLY THIS ITEM'S CTF

Defocus: 1,190 µm
Astigmatism: 0.010 µm, 81.0°
Phase shift: 0.000 m
Res. estimator: 2.1 Å

Zoom: 0.25 x Intensity range: 2.50 e-
☒ Deconvolve RETRAIN ON THIS DATA SET
☒ Show motion tracks, 20 x scale, 8 x 8 grid ☐ only local motion ☒ Show defocus, 1,181 µm — 1,200 µm
☒ Show particles from [Boxer/Mask_20180915](#) with 100 Å diameter, at least 0.000 score ☒ Dots ☐ flash — 114 particles
[PICK WITH BOXNET2MASK20180915](#) ☐ Show mask, PAINT with a 300 Å brush

EMD-2084_0012_Frames (13/1539)

Figure 2.2 | User interface of *Warp*.

a) The processing settings (left) specify all steps and parameters for online data evaluation, correction and processing. The 'Overview' tab (right) presents all important processing results and lets the user specify selection filters to remove low-quality data.

b) View of a single micrograph. In Fourier space (left), the simulated 2D CTF (i), the 1D power spectrum (PS) and its fit (ii), and the 2D PS (iii) are presented. The real space view (right) shows the aligned movie average with particle positions (green dots), motion tracks (white curves) and the defocus variation (transparent magenta-cyan overlay), and applies a deconvolution filter as well as denoising. Individual display elements can be shown or hidden. The navigation bar (bottom) shows the processing status for all items and allows to quickly switch between them as well as to manually exclude single items from processing.

motion, or particle count, for all processed micrographs or tilt series as interactive plots. The user can immediately grasp the statistical distribution, observe intrinsic patterns, and make an informed decision to adjust the acquisition strategy, e.g. to tune the lens astigmatism, increase the stage settling time, or skip a bad grid square. A filter range can be specified for every plotted parameter to automatically exclude lower-quality images from downstream processing. Any data point can be quickly inspected in more detail in a tab called 'Fourier & Real Space'. Here, a display of the power spectrum and the CTF fit allows optimization of CTF fitting parameters. In the real-space view, a deconvolution filter and neural net-based denoising (Methods, **Figure 2.3**) can be instantly applied to micrographs to improve the contrast and make the particles more visible to the human eye. The defocus variation obtained through local CTF estimation can be overlaid semi-transparently. Particle picking results can be assessed in the context of a micrograph, and edited manually. Dedicated dialogs assist the user with tasks like micrograph list export, particle extraction, template matching, tomogram reconstruction, and neural network training.

2.1.4 Motion correction

Warp generally represents space- and time-dependent parameters as coarse, uniform grids, on which a computationally cheap B-spline interpolation can retrieve any intermediate value (Methods). The motion, i.e. the translational shift observed between frames,

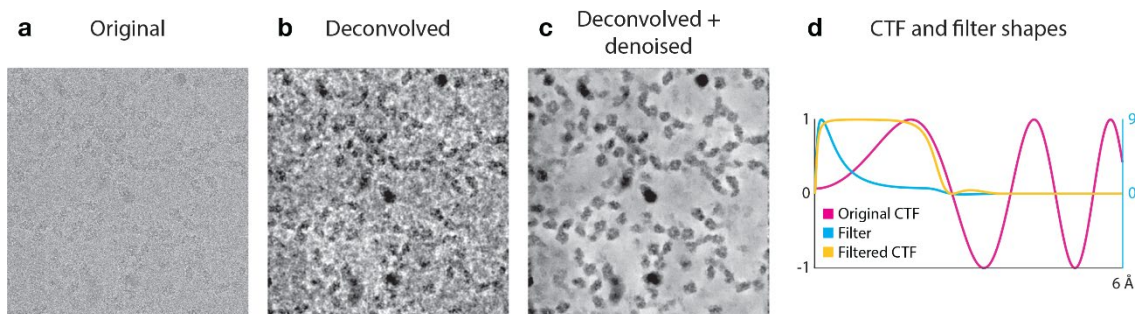


Figure 2.3 | Deconvolution and denoising of a low-defocus micrograph.

- a)** A raw micrograph from EMPIAR-10061 acquired at 0.8 μm defocus.
- b)** Same micrograph after applying deconvolution. Low-resolution contrast is boosted and the defocused signal is more localized, allowing to distinguish the particles better.
- c)** Same micrograph after applying deconvolution and denoising with a noise2noise model retrained on this data set. The shapes of individual 400 kDa proteins nearly invisible in the raw image can be distinguished clearly against the background.
- d)** Shape and effect of the deconvolution filter. The filter largely reverses the effect of the first CTF peak, while also suppressing the lowest and higher frequencies.

is due to two effects: movement of the mechanical sample stage, and BIM. Stage movement causes global shifts over the entire field of view, whereas BIM leads to shifts between adjacent micrograph patches^{10, 21}. While stage drift can lead to rapid shift changes between frames, BIM occurs more slowly after rapid relaxation during initial exposure⁹. *Warp* corrects for both global drift and local BIM at variable temporal resolution (**Figure 2.4**). The strategy is similar to the one used by MotionCor2²³, except that *Warp* does not apply additional *a priori* assumptions about BIM beyond those imposed by the parameter grid resolution. As a result, *Warp* corrects very efficiently and thoroughly for the two types of motion that occur during cryo-EM data acquisition in any kind of support film hole morphology and orientation.

2.1.5 Estimation of local defocus and resolution

The CTF model can be estimated based on the power spectrum (PS) of a micrograph. However, the defocus varies over the micrograph area due to stage inclination, uneven sample surface, or an uneven particle distribution along the optical axis¹⁵. *Warp* provides a flexible way to model local defocus variation in spatial and temporal dimensions without the need for *a priori* knowledge of particle positions. Instead of one global estimate,

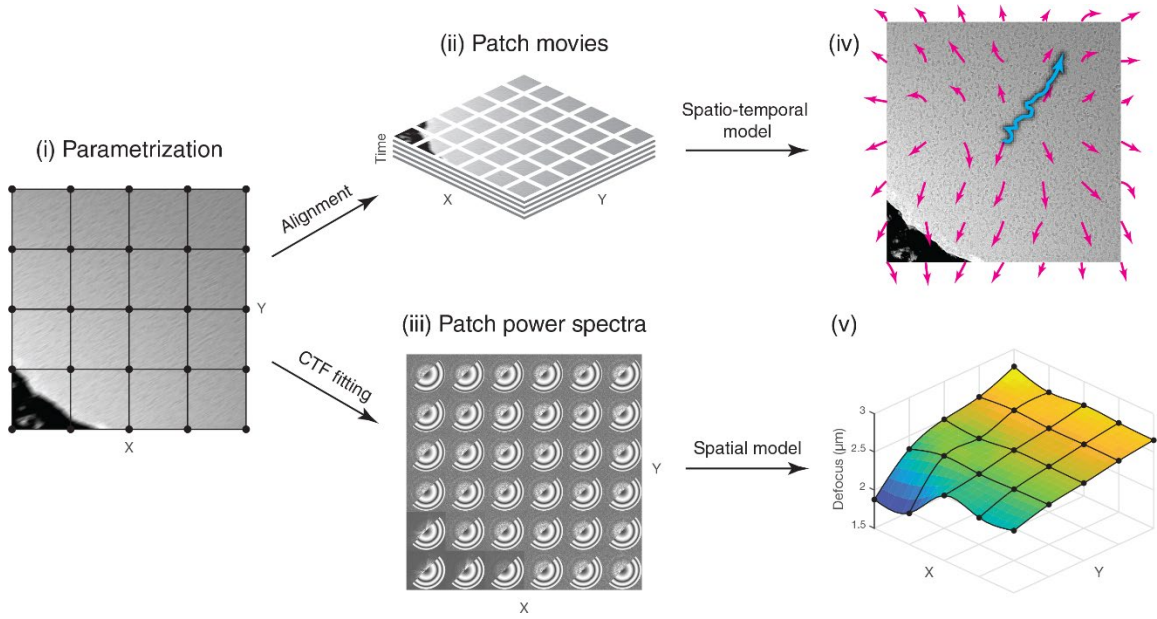


Figure 2.4 | Motion and CTF model fitting by Warp.

The unaligned, defocused movie (i) is parametrized with a coarse grid (black dots), divided into patches for the alignment (ii), and power spectra of these patches are computed (iii) for CTF fitting. The motion model (iv) includes 2 components: global motion (cyan trajectory) with fine temporal and no spatial resolution, and local motion (magenta trajectories) with coarse temporal, and fine spatial resolution. Both components are optimized to minimize the squared difference between the individual patch frames and their aligned average. The spatially resolved CTF model (v) is optimized to minimize the squared difference between the power spectra (iii, upper left part of each patch) and the simulated local 2D CTF (iii, bottom right part of each patch). Here, the defocus gradient follows the 40° tilt of the specimen, with the notable exception of the hole edge in the bottom left corner.

a tilted plane or a more complex geometry is fitted to the PS of a movie patch to estimate local defocus. A 1D average of all local power spectra rescaled to a common defocus value allows the user to easily assess whether fitting the more complex geometry recovered more Thon rings beyond the spatial frequencies used for the fitting (**Figure 2.5**). Thus, *Warp* goes beyond state-of-the-art CTF estimation by providing a spatially resolved model without the need for *a priori* knowledge of particle positions, and costly hyperparameter tuning. The spatially resolved CTF model can converge on the correct solution for tilts as high as 60°, based on the inclination of the estimated defocus gradient. This makes *Warp* a useful tool for tilted 2D data collection, which has been shown to increase

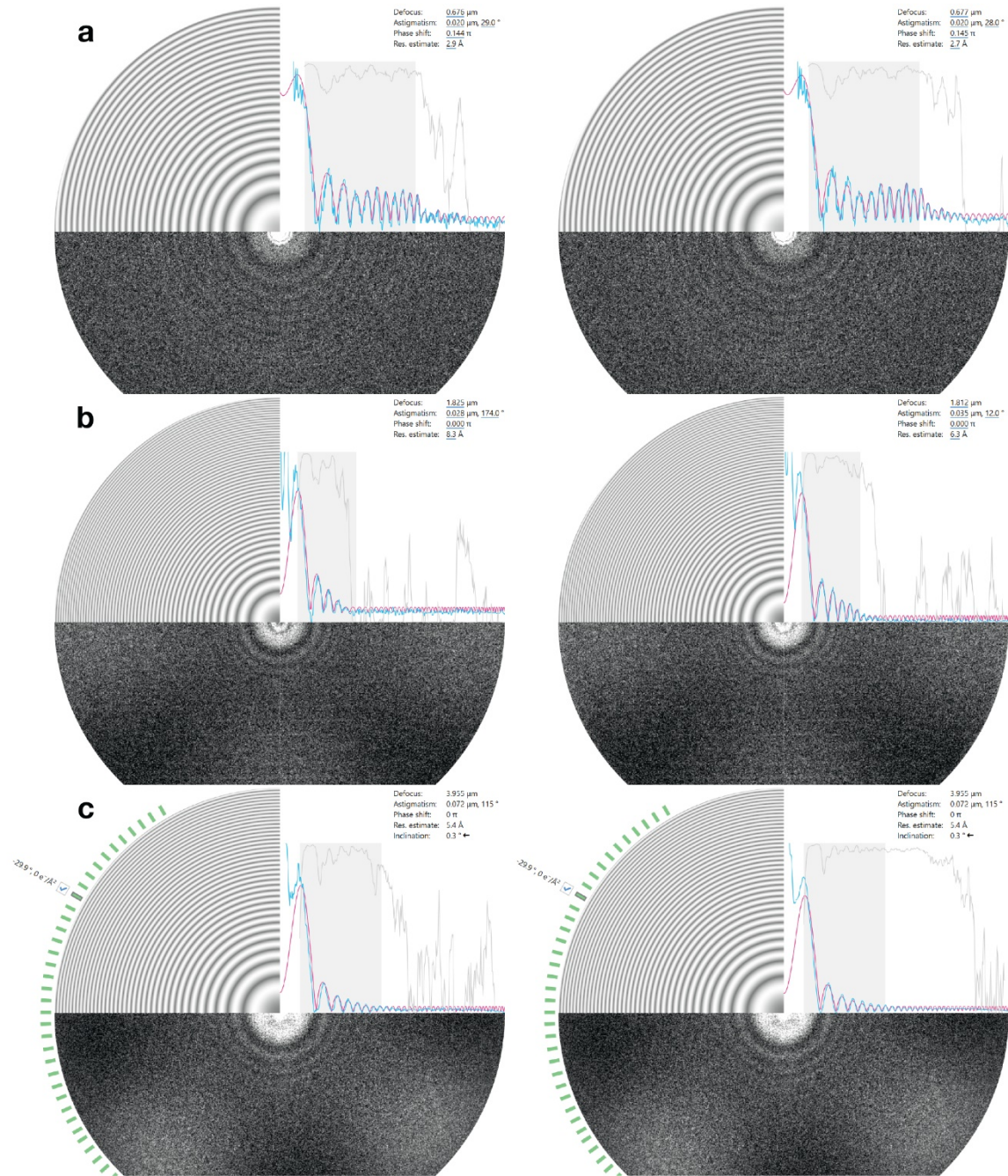


Figure 2.5 | CTF fitting of flat, tilted and tilt series data.

Fitted spectra without (left column) and with (right column) a spatially resolved defocus model. The samples are **(a)** flat (EMPIAR-10078), **(b)** tilted at 40° (EMPIAR-10097), **(c)** a tilt series ranging from -60° to +60° (EMPIAR-10045). In all three cases, using a spatially resolved model allowed to fit the sample geometry more accurately, as evidenced by the clearer Thon rings in the rescaled, averaged 1D spectra. The fitting range (grey rectangle in the 1D spectra) was chosen well below the estimated resolution to avoid overfitting the higher number of parameters in the spatially resolved model.

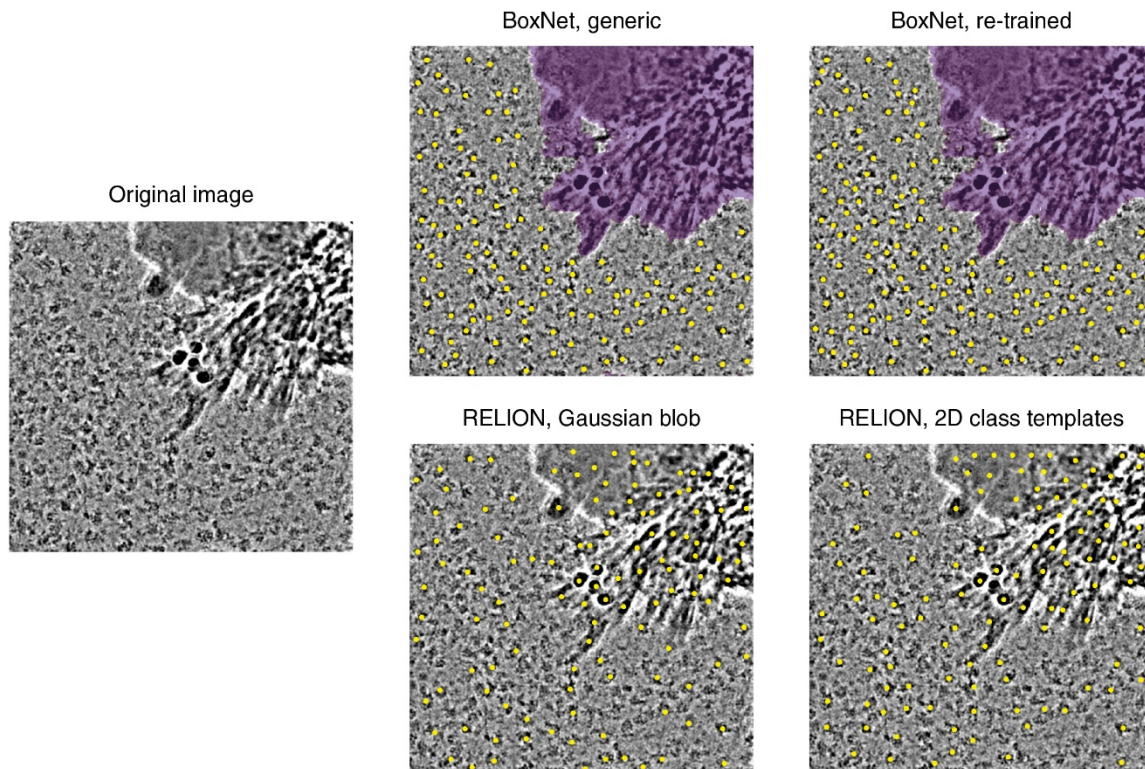


Figure 2.6 | Automated particle picking with Warp's deep learning-based BoxNet.

Representative example of automated particle picking with BoxNet in Warp on a micrograph with high-contrast artifacts. Areas masked out automatically by BoxNet are colored purple. The generic version of BoxNet was never presented with the sample during training. The re-trained version was given 5 micrographs of the same sample, which did not include the one shown. The template-based picking in RELION used 25 class averages derived from 3000 particles, filtered to 20 Å. RELION's results show the 120 highest-scoring positions. For visualization purposes, the micrograph was deconvolved, high-pass filtered and cropped at the borders.

the resolution isotropy for samples with preferred orientation³⁰. The useful resolution range of a micrograph is estimated as the spatial frequency where the fit quality falls below a threshold (Methods).

2.1.6 Particle picking with BoxNet

The next step in cryo-EM structure determination is the accurate selection of single particles from the corrected micrographs. *Warp* includes a novel particle picking routine based on a machine learning algorithm (Methods). For several years, the computer vision community has been using convolutional neural nets (ConvNets) to vastly outperform

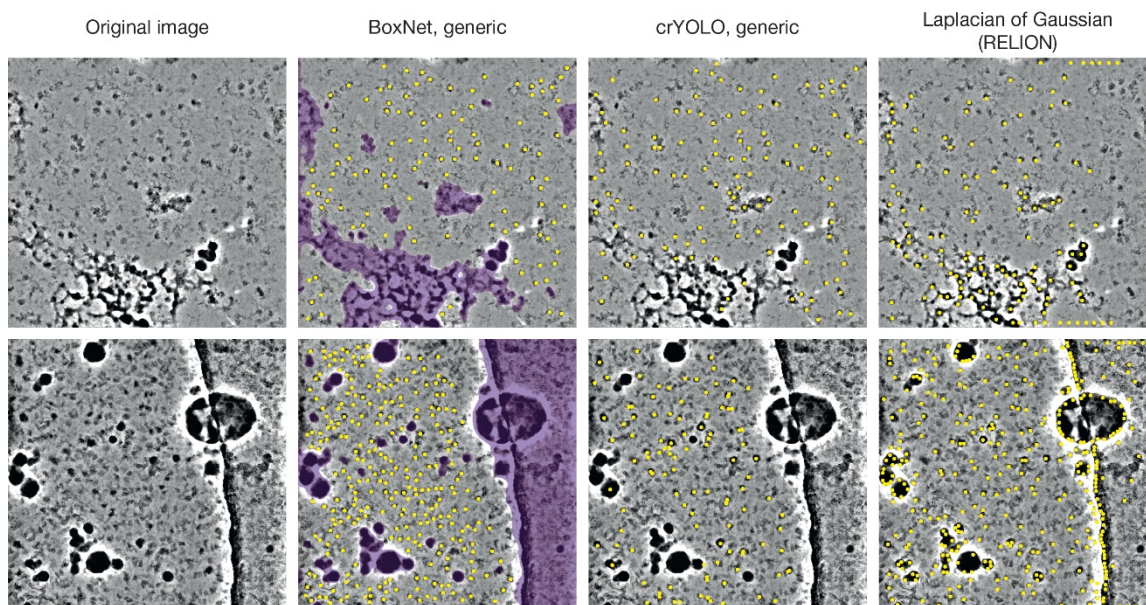


Figure 2.7 | Unbiased particle picking with Warp's BoxNet.

Examples of automated particle picking on samples not seen by BoxNet in training. For comparison, the same micrographs were picked with crYOLO's generic model, and RELION's Laplacian of Gaussian (LoG) method. Micrographs were selected from in-house data to make sure they were absent in crYOLO's knowledge base. BoxNet reliably recognizes almost all particles (yellow), and masks out all artifacts (purple). LoG is often confused by high-contrast edges and ethane impurities. crYOLO performs better than LoG, but is also routinely confused by ethane impurities and protein aggregates, and misses many of the small particles (bottom row).

template matching in object recognition tasks^{54, 55}. First attempts to apply ConvNets to the particle picking problem in cryo-EM have shown performance on par with traditional template matching approaches⁵⁶. Today, deep residual network (ResNet) architectures enable the training of arbitrarily deep models⁵¹. *Warp* employs 'BoxNet' – a fully convolutional ResNet architecture with 72 layers, implemented in TensorFlow 1.5⁵⁷. BoxNet was trained with data from the EMPIAR raw data repository⁵⁸ and synthetic data simulated from PDB⁵⁹ structures with a molecular weight range of 0.064–18 MDa. As a result of these efforts, the pre-trained neural network bundled with *Warp* performs well on many particle species and is able to accurately mask out high-contrast artifacts, such as ethane. The performance of BoxNet compares favorably with available tools when representative single-particle cryo-EM data are used as input (**Figure 2.6**). For heterogeneous

data sets, BoxNet’s generic model can provide an advantage compared to the generic model in crYOLO⁶⁰ or the Laplacian of Gaussian approach in RELION 3.0³³ (**Figure 2.7**).

2.1.7 Retraining of BoxNet

Since the performance of BoxNet can vary between different data, *Warp* offers a retraining interface for BoxNet. Such retraining leads to a very high accuracy in automated particle picking. For retraining, the user can indicate to *Warp* positive and negative examples of BoxNet performance. Using ~1000 examples, retraining of BoxNet typically takes less than 10 minutes, with an estimate of the achieved accuracy provided during the process. After retraining, the user can pick the same micrographs with the re-trained network and select more positive and negative examples for another round of retraining if required. To decrease the need for retraining in the future, *Warp* also provides the option of submitting training data to a central GitHub repository. *De novo* training will be carried out by us periodically with all deposited data, and the resulting updated pre-trained BoxNet offered to the community. The training set is centrally curated and a list of particle species in the current version is available from <https://github.com/cramerlab/boxnet>. The BoxNet version name will be stored in each micrograph’s metadata to ensure reproducibility of picking results obtained with older versions.

2.1.8 Online pre-processing during data collection

The design of *Warp* is optimized for processing raw cryo-EM data immediately during data collection. Files written out by the image acquisition software are detected automatically in the specified input folder and added to the list of ‘processable items’ in *Warp*. Each item maintains its metadata in an XML file that includes the exact previous processing settings. *Warp* continuously performs the processing steps necessary to bring each item into accord with the settings currently specified for the entire folder. All results can be immediately inspected during processing. Items can be forcibly included (i. e. exempted from the quality filters) or excluded from downstream processing. The processing must be stopped to change the settings or to retrain the BoxNet model. If changes were made, *Warp* will first reprocess all outdated items. During online processing, *Warp* is able

to estimate parameters such as motion, defocus and the resolution limit from micrographs, as well as perform particle picking within less than one minute after the raw data become available. In our experience, high-quality single-particle data of complexes of RNA polymerase II enable the user to obtain detailed 2D classes of particles and 3D reconstructions at better than 5 Å resolution using *Warp* and cryoSPARC within only a few hours after the start of data collection.

2.1.9 Interoperability with other software

To ensure interoperability with a plethora of cryo-EM tools, *Warp* allows the user to import and export data at several steps in its workflow using widely accepted formats and standards. Raw movie data in the MRC, TIFF and EM formats are supported, and a ‘headerless’ option allows the user to manually specify properties of an unknown binary format. Data are exported in the widely used MRC format, whereas all metadata are saved in the STAR format, adhering to the conventions established by RELION and adopted by many other tools. All pre-processing steps can be turned off if required. Results obtained with other tools can be imported to skip or benefit from particular algorithms in *Warp*. For instance, particle positions can be imported to export aligned particle averages, update their CTF models with *Warp*’s local estimates, to obtain a comprehensive overview of the particle distribution in a large project, or to retrain a BoxNet model. Frame alignment data can be exported to initiate a more accurate, reference-based alignment in RELION 3.0. Micrograph and particle lists adhering to user-selected quality criteria can be quickly prepared and exported. Taken together, *Warp* is highly flexible and allows for easy interoperation with other cryo-EM data processing tools used by the community.

2.1.10 Pre-processing tomographic data

Warp can also be used to pre-process data from cryo-electron tomographic (cryo-ET) tilt series. *Warp* can reconstruct tomograms from a tilt series and can perform template matching in tomograms with available 3D structures. The (sub)-tomogram reconstruction considers the local CTF, sample distortion and magnification anisotropy (Methods). Additionally, a deconvolved version of the tomograms can be produced using the same

interface to help with their visual evaluation. To ensure the CTF model is as accurate as possible, *Warp*'s CTF fitting procedure goes beyond fitting the tilt images individually. Instead, local patch 2D power spectra from all tilts are fitted simultaneously, with constraints imposed on the inter-tilt angles, and regularizing assumptions made for the progression of phase shift and astigmatism (Methods). The tilt series CTF fitting can also be performed as part of the online processing.

2.1.11 Template matching

In addition to picking particles of a new species of unknown shape with a system like BoxNet, finding a previously known structure in new data is central to many stages of cryo-EM data processing. The structure must be compared at many different orientations to every position in the new data under the consideration of the CTF. *Warp* implements template matching only for 3D templates, because matching a set of *de novo* 2D templates for particle picking is better handled by a neural network such as BoxNet (see above). A template volume can be either provided by the user or automatically downloaded from the EMDB⁶¹ through the same UI. For template matching, 2D micrographs are subdivided into tiles. Then, normalized cross-correlation is computed between the tiles and 2D projections generated from the 3D template at the specified angular intervals, convolved with the local 2D CTF (Methods). All local correlation peaks with a minimum inter-peak distance corresponding to the template particle diameter, and the corresponding best-scoring template orientations are saved so that the user can later instantly explore different peak thresholds without repeating the costly correlation step. This procedure is also implemented for tomographic volumes, where the local patches are replaced by local sub-volumes, and the local 3D CTF is considered to incorporate knowledge of defocus and the missing wedge. In both cases, the matching should be performed at a resolution significantly lower than the expected map resolution to avoid template bias⁶². In case the map resolution does not surpass the resolution used for template matching, the procedure should be rerun at a lower resolution, and the results reprocessed.

2.1.12 Software implementation

Warp is written in the programming languages C#, C++ and CUDA C. The expressiveness of C# and the availability of powerful development tools kept the high-level data management layer brief and maintainable. *Warp*'s rich UI is enabled by the Windows Presentation Foundation (WPF) framework. All performance-critical parts are implemented to run on a GPU. Central data primitives, such as 2D movies and tilt series, and all associated algorithms are wrapped in a stand-alone C# library that we called 'WarpLib'. The granularity of most of these methods is fine enough to make them useful for applications beyond those implemented in *Warp*. Thus, WarpLib has the potential to speed up the development of future GPU-enabled tools that provide new functionality around the same data. We intend to keep developing *Warp* to enable state-of-the-art, rapid cryo-EM data pre-processing in the future. Updates will be shared with the community via GitHub.

2.1.13 Benchmarking for 2D data

To test the performance of *Warp*, we reprocessed a published single-particle cryo-EM data set for the influenza hemagglutinin trimer³⁰ (Methods) (**Figure 2.8**). We chose this case for benchmarking because the processing of a 150 kDa protein imaged at 40° tilt required a significant amount of manual screening in the original analysis³⁰, providing a challenging test case for the *Warp* pipeline. With the original set of 130,000 particles, cryoSPARC reached a similar resolution as that reported in the original analysis (**Figure 2.8a, b**), showing that refinement in cryoSPARC and RELION yields equivalent results for this data set. However, because this particle set and the general particle population both exhibit significant heterogeneity, we draw the comparison between results obtained after subjecting all data to the same 3D classification steps in cryoSPARC (Methods). For the original set, the best class containing 57,346 particles reached a global resolution of 3.9 Å with a B-factor of -200 Å². The same particles, updated with the defocus information from *Warp*, reached a notably higher resolution of 3.5 Å with a B-factor of -170 Å². This suggests that *Warp*'s local CTF model is more accurate than the per-particle CTF fitting in gCTF²⁹ used in the original study. *Warp* processing also estimated a narrower range of

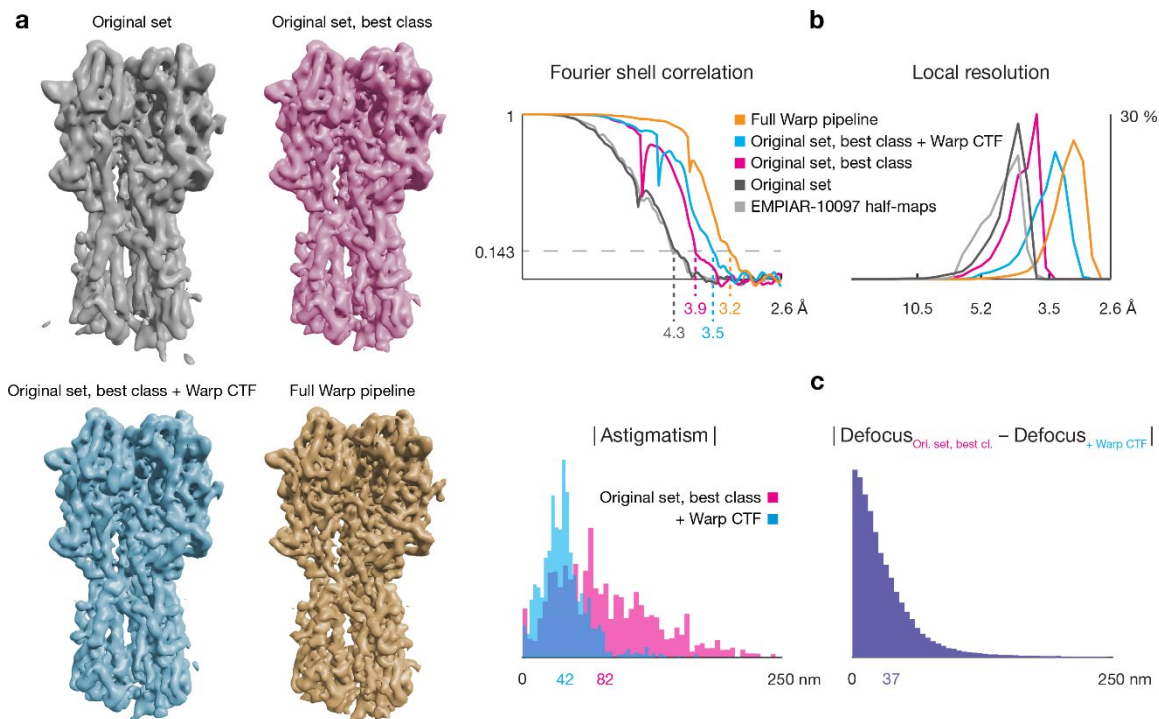


Figure 2.8 | Warp's 2D pipeline improves cryo-EM density for influenza hemagglutinin.

As a benchmarking case we used the published EMPIAR-10097 data set containing influenza hemagglutinin trimer particles. 'Original set': 130,000 pre-extracted particles from EMPIAR-10097 with their original CTF parameters; 'Original set, best class': 57,346 particles from 'Original set' after 3D classification in cryoSPARC with their original CTF parameters; 'Original set, best class + Warp CTF': the same 57 346 particles, but with Warp's CTF estimates; 'Full Warp pipeline': 249,495 particles obtained from the raw EMPIAR-10097 data after unsupervised pre-processing in Warp and 3D classification in cryoSPARC.

a) Isosurface renderings of the 3D maps generated with cryoSPARC using the respective sets of particles and CTF parameters, filtered to local resolution using the auto-tightened masks generated by cryoSPARC.

b) Global masked FSC plots, and histograms of the local resolution used to filter the maps depicted in (a). 'EMPIAR-10097 half-maps' refers to the original half-map volumes deposited in EMPIAR-10097, obtained from the same 130,000 particles as 'Original set'.

c) Histogram comparison between the original defocus parameters and those estimated by Warp for the 130 000 particles from 'Original set'. The mean value for each metric is specified underneath the horizontal axis in the same color as the corresponding histogram.

astigmatism amplitudes (**Figure 2.8c**), in agreement with the assumption of a stable optical system. For the full, completely automated *Warp* pre-processing pipeline, the best class containing 249,495 particles reached a global resolution of 3.2 Å with a B-factor of -

170 Å², accompanied by a significantly increased level of detail in the map (**Figure 2.8a**). This improvement from 3.5 Å to 3.2 Å is due to the higher particle count, which was obtained by *Warp* in a fully automated fashion at no time cost to the user. After classification in cryoSPARC, the best classes contained 45% and 51% of all particles in the original EMPIAR-10097 set and *Warp*'s automatically picked set, respectively, suggesting a similar degree of particle 'purity' in the manual and automated approaches. Furthermore, this demonstrates that tilted data collection can lead to near-atomic resolution with minimal efforts at the data pre-processing step.

The pre-processing of 668 movies from EMPIAR-10097 in *Warp* required ca. 3 h using a system with 4 Nvidia Titan X GPUs, thus averaging to ca. 220 movies per hour. This speed allows *Warp* to keep up with any of the current automated acquisition schemes, which collect ca. 90 movies per hour on a Titan Krios microscope (Thermo Scientific, USA), and can reach up to 150 movies per hour in the most favorable cases (Wim Hagen, EMBL – personal communication). Taken together, our results establish *Warp* as a very useful tool for high-performance, automated cryo-EM data pre-processing.

2.1.14 Complementarity of *Warp* with other tools

We further tested *Warp*'s performance on a data set of β-galactosidase particles⁶³ that is often used for benchmarking purposes (**Figure 2.9**). Because of the sample's high structural stability, these data stress the software's ability to obtain particularly accurate CTF and motion estimates to reach the highest end of cryo-EM resolution. Furthermore, the particles are difficult to pick due to the low defocus and prevalence of high-contrast objects in the micrographs. To estimate the frame alignment accuracy independently of 3D refinement, we calculated the average CTF fit quality for aligned movie averages processed with MotionCor2²³ or *Warp*. *Warp*'s averages could be fitted to 2.6 Å, and those of MotionCor2 to 2.7 Å (**Figure 2.9c**), indicating slightly better frame alignment in *Warp*. After refinement of particles from the full, completely automated *Warp* pre-processing pipeline, the best class containing 127,000 particles reached a global resolution of 2.09 Å with a B-factor of -35 Å².

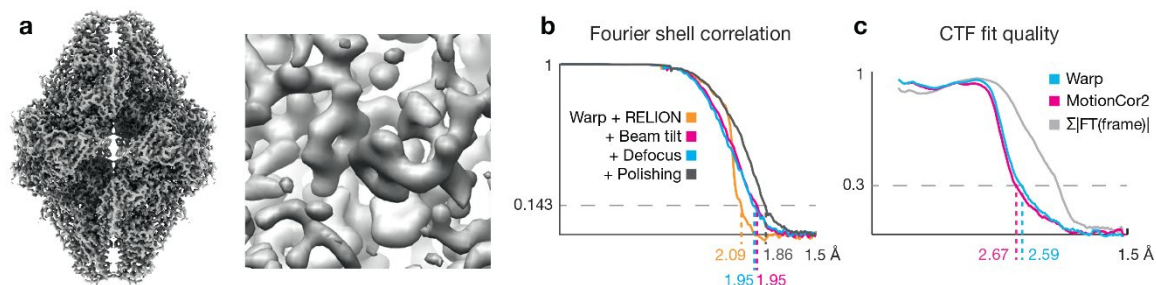


Figure 2.9 | Warp’s 2D pipeline in combination with RELION 3.0 improves cryo-EM density for β -galactosidase.

For our second benchmark, we used the published EMPIAR-10061 data set containing β -galactosidase particles. The data were processed using the full *Warp* pre-processing pipeline, and beam tilt, per-particle defocus and frame alignment were later refined against high-resolution references in RELION 3.0 to assess the additional improvement provided by these refinements.

a) Isosurface rendering of the 1.86 Å map (left) and a detailed view of some of its sidechains, clearly displaying the aromatic rings (right).

b) Global masked FSC plots for the map obtained with the Warp pipeline only, and for the additive effects of reference-based beam tilt and per-particle defocus refinement, as well as particle polishing in RELION 3.0.

c) Average CTF fit quality curves for aligned movie averages produced with MotionCor2 and Warp. Warp’s averages can be fitted to a higher resolution, indicating more accurate frame alignment. For comparison, a fit quality curve is also included for amplitude spectra obtained from the average of individual frame spectra, which are invariant to residual inter-frame motion and radiation damage.

To test the advantage of using reference-based refinements implemented in RELION 3.0³³, the same particle set was first subjected to beam tilt refinement and reached a resolution of 1.95 Å with a B-factor of -29 Å^2 . Adding per-particle defocus refinement did not improve the resolution further. Adding reference-based frame alignment, referred to as ‘polishing’ in RELION, improved the resolution further to 1.86 Å with a B-factor of -26 Å^2 . This slightly surpasses the result reported³³ for the same set of refinements in RELION 3.0 by 0.04 Å, possibly through cleaner initial particle picking. Whereas the additional refinements improved the resolution beyond that achieved with the pure *Warp* pre-processing pipeline, we note that these procedures required the data to be fully classified and refined first. In contrast, *Warp* provided its accurate initial results before any downstream processing took place. This was evident for the hemagglutinin trimer data, where

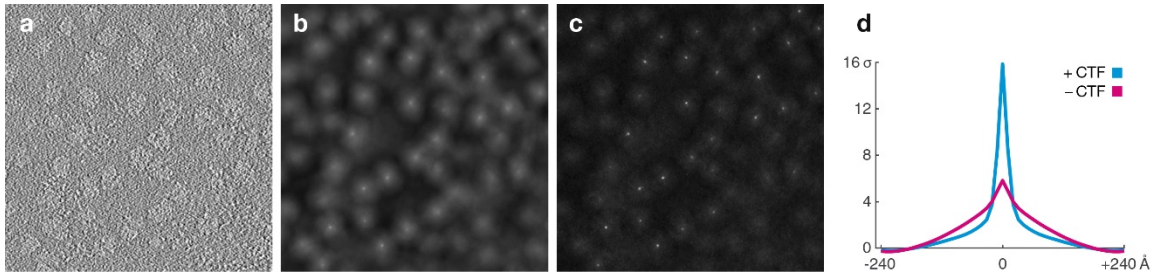


Figure 2.10 | Effect of using the full local 3D CTF for template matching in tomograms.

During template matching, *Warp* multiplies the rotated 3D reference by the local 3D CTF before correlating it to a local portion of the tomogram volume, as opposed to multiplying it by a binary missing wedge mask. This produces sharper correlation peaks.

a) XY slice through a tomogram reconstructed from EMPIAR-10045 data. The faint shapes of 80S ribosomes are visible.

b) XY slice through the correlation volume at the same location as (a), using a binary missing wedge mask. White indicates higher correlation. The peaks are broad and hard to distinguish against the background.

c) XY slice through the correlation volume at the same location as (a), using the full local 3D CTF. White indicates higher correlation. The peaks are sharper, leading to higher template matching accuracy.

d) Rotational average of a 48 px window around all correlation peaks, mean-subtracted and normalized against the respective correlation background. 3D CTF-aware template matching (+CTF) produces peaks rising 2.7 times higher above the background compared to binary missing wedge masks (−CTF).

Warp obtained the local defocus values immediately, whereas several rounds of refinement were required by RELION 3.0³³. Thus, the pre-processing in *Warp* and reference-based refinements in RELION 3.0 are complementary approaches, whose combination leads to faster convergence and higher resolution.

2.1.15 Benchmarking for tilt series data

To assess the benefits of using the full local 3D CTF for template matching in tomograms compared to a binary missing wedge mask, we matched 7 tomograms from a publicly available 80S ribosome data set⁶⁴ with the 3D map published as a result of its initial processing (**Figure 2.10**). Among the 3,288 top-scoring matches, the false positive rate was 1% when using the full local 3D CTF, and 15% when using the binary missing wedge mask. At 15.9 standard deviations, the CTF-aware correlation peaks rose 2.7 times higher above

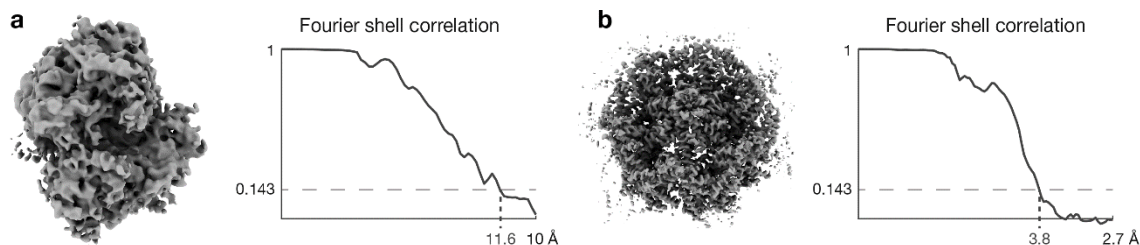


Figure 2.11 | Sub-tomogram averaging results obtained by using *Warp*'s tilt series CTF estimation and sub-tomogram export.

To assess the benefits of the proposed CTF estimation and sub-tomogram export strategies, data from EMPIAR-10045 and EMPIAR-10164 were pre-processed and exported in *Warp*, and refined in RELION 3.0. Improved resolution was observed in both cases compared to published results.

a) Isosurface rendering (left) and FSC plot (right) of the 80S ribosome sub-tomogram average obtained from EMPIAR-10045 data. The originally published resolution was 12.8 Å. **b)** Isosurface rendering (left) and FSC plot (right) of the HIV-1 sub-tomogram average obtained from 12% of the EMPIAR-10164 data. The originally published resolution for this subset was 3.9 Å.

the respective background distribution than the 5.8 standard deviations of the CTF-unaware peaks (**Figure 2.10**), thus allowing more accurate template matching.

To benchmark the proposed CTF estimation routine and sub-tomogram export for tilt series, we pre-processed, exported and refined particles from two publicly available data sets. For EMPIAR-10045 and EMPIAR-10164 (subset of 5 tomograms used in the assessment of NovaCTF⁶⁵), we obtained a resolution of 11.6 Å with 3,200 particles (**Figure 2.11a, b**), and 3.8 Å with 22,000 particles (**Figure 2.11c, d**), respectively. These results slightly surpass the resolution figures of 12.8 Å and 3.9 Å reported in the original studies. Thus, the new tilt series processing algorithms implemented in *Warp* can improve upon the state of the art, leading to higher resolution in sub-tomogram averaging.

2.2 Methods

2.2.1 Spline interpolation on multi-dimensional grids

Many methods in *Warp* are based on a continuous parametrization of 1—3-dimensional spaces. This parameterization is achieved by spline interpolation between points on a coarse, uniform grid, which is computationally efficient. A grid extends over the entirety

of each dimension that needs to be modeled. The grid resolution is defined by the number of control points in each dimension and is scaled according to physical constraints (e. g. number of frames or pixels) and available signal. The latter provides regularization to prevent overfitting of sparse data with too many parameters. When a parameter described by the grid is retrieved for a point in space (and time), e. g. for a particle (frame), B-spline interpolation is performed at that point on the grid. To fit a grid's parameters, in general, a cost function associated with the interpolants at specific positions on the grid is optimized. In the following, we distinguish between 2—3 spatial grid dimensions (X and Y axes in micrographs; X, Y and Z in tomographic volumes), and a temporal dimension as a function of the accumulated electron dose. We note that B-splines are only used to interpolate parameters, not image data. For the latter higher-order schemes are used.

2.2.2 Motion model

Two sources contribute to the observed translational shift between frames in a dose-fractionated image sequence. First, mechanical stage instability leads to rapid shift changes that are uniform within the entire frame. Second, beam-induced motion (BIM) causes slowly changing, local motion. Warp considers the physical properties of both sources in its motion model, using two sets of grids to parametrize the frame shifts and sample deformation. Global motion is described by two grids, X_{global} and Y_{global} , with high temporal, and no spatial resolution. The temporal resolution can match the number of frames, or, in case finer dose fractionation is performed to reduce intra-frame motion, the resolution can be lower to regularize a potentially overfitted model. BIM is described by two grids, X_{local} and Y_{local} , with a temporal resolution of at most 3, and a spatial resolution of typically 4—5 in both dimensions. The overall shifts required to bring the same object in all frames into a common reference frame are then defined as $(X_{\text{global}} + X_{\text{local}}, Y_{\text{global}} + Y_{\text{local}})$.

2.2.3 Global and local motion correction

In the absence of known particle positions and high-resolution reference projections, individual frame patches are aligned to their averages. The movie is subdivided into groups

of 512^2 px patches with a 50 % spatial overlap, masked with a raised cosine. To simplify computation, the images are transformed into Fourier space where complex multiplication replaces translation. For each group, the patches are shifted according to the interpolants at their extraction positions using the current grid values. The average of a group's shifted patches is then compared to the individual patches to calculate the patch costs as

$$C = \sum_i \sum_f |I_{i,f} - \bar{I}_f|^2,$$

where i denotes the frame index, f denotes the spatial frequency, I is the Fourier transform of a shifted patch frame, and \bar{I} is the average of all shifted patch frames. The shifts are obtained by interpolating on the current state of the parameter grids at the patch frame's position in space and time. The derivative is approximated numerically with the symmetric difference quotient. The overall cost for all grid control points is the sum of all patch costs, and the derivative for each grid control point is the weighted sum of the derivatives of all patches affected by it. The weights for each control point's derivative can be precomputed by applying a one-pixel shift to the control point and storing all resulting non-zero patch shifts. The cost and derivatives are used by the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm⁶⁶ to optimize the values of all control points. The optimization is performed in several steps to improve global convergence. In the first step, the temporal resolution of X_{global} and Y_{global} is set to 3, and increased to the next power of 3 in subsequent steps until the desired temporal resolution is reached.

2.2.4 Contrast transfer function estimation in micrographs

The CTF analytically describes the convolution applied to the images by the electron-optical system. Estimating its properties with high precision is essential for reversing the effects and obtaining high-resolution reconstructions⁶⁷. Whereas the methodology for measuring defocus and astigmatism from a micrograph's power spectrum (PS) has been well-established^{27, 68}, the recent increase in EM map resolution calls for a more localized approach. Local defocus variation of a seemingly flat sample can exceed 60 nm within a

single micrograph, resulting in an out-of-phase CTF for some particles at resolutions beyond 3 Å. Attempts to address this issue by fitting the defocus per-particle have been made²⁹, but they require knowledge of particle positions, and lack robustness for all but the largest particle species. Even with a local smoothing approach, per-particle defocus requires high particle density to not lose accuracy compared to a global estimate. On the other hand, strong local irregularities in the specimen surface are almost never observed in tomographic volumes in vitro¹⁵, suggesting per-particle precision might be unnecessary.

2.2.5 Estimation of local defocus

To parametrize the defocus, a single grid typically consists of 5x5 spatial control points and 1 temporal control point. Optionally, another single grid with exclusively temporal resolution tracks the development of the phase shift generated by a Volta Phase Plate¹². Global parameters, such as astigmatism magnitude and angle, are optimized as additional scalars in the model. In practice, the effect of using a more complex geometry or temporal resolution appears negligible. However, the increased signal of future camera hardware might make these options relevant. Like in the frame alignment procedure, groups of patches matching the desired PS size (e. g. 512^2 — 1024^2 px) are extracted with a spatial overlap of 50% from the raw movie data, transformed into Fourier space, and converted to PS by taking their squared amplitudes. If no temporal resolution is desired, each group will be averaged to a single PS to save resources. Similarly, in the absence of spatial resolution, the same frame from all groups will be averaged.

For the initial, exhaustive search, a 1D rotationally averaged PS is calculated from all patches. A B-spline with 3 control points is then fitted through it and subtracted to remove most of the background. The user-defined range of defocus and, optionally, phase shift values is evaluated by matching a simulated 1D CTF. The result with the highest normalized correlation is then used to estimate the 1D PS background and envelope more accurately to consider both in subsequent 2D CTF fitting. The cost function to be

minimized is calculated between a 2D PS, and the 2D CTF simulated based on the local defocus at the patch extraction position:

$$C = \sum_i \sum_f (|CTF_{i,f}| \cdot E_f)_{Norm} \cdot (|I_{i,f}| - B_f)_{Norm},$$

where i denotes the frame index (in case a temporal dimension is used), f denotes the spatial frequency within the range used for fitting, CTF is the 2D contrast transfer function, I is the FT of a patch frame, E is the envelope of the 1D PS, B is the background of the 1D PS, and $(\dots)_{Norm}$ is a normalization operator that brings the value distribution to mean = 0, standard deviation = 1. The defocus and, optionally, phase shift is obtained by interpolating on the current state of the respective parameter grid at the patch frame's position in space and time. The derivative is approximated numerically with the symmetric difference quotient. The same pre-computed weights strategy as in the frame alignment procedure is employed for the control point derivatives. An L-BFGS algorithm finally optimizes the model for all control point values.

The PS of a tilted plane will usually only show low-resolution Thon rings, regardless of what model was used for the defocus gradient. To provide the user with feedback on whether the more complex defocus model is beneficial, the 2D spectra from all patches, whose parameters are herein referred to as the “source” PS, are rescaled and rotationally averaged to a 1D PS with a single defocus value, referred to as the “target” PS, such that the CTF phases match using the following scaling function:

$$x'_\varphi = \sqrt{\left| \frac{-\sqrt{p_\varphi'^4 \cdot p_\varphi^4 \cdot (C_S^2 \cdot \lambda^4 \cdot f_N^4 \cdot x_\varphi^4 + 2C_S \cdot \lambda^2 \cdot f_N^2 \cdot p_\varphi^2 \cdot x_\varphi^2 \cdot z_\varphi + p_\varphi^4 \cdot z_\varphi'^2)} - p_\varphi'^2 \cdot p_\varphi^4 \cdot |z'_\varphi|}{C_S \cdot \lambda^2 \cdot f_N^2 \cdot p_\varphi^4} \right|},$$

where $'$ denotes the “target” PS coordinate system, and its absence denotes the “source” PS coordinate system; φ is the sampling angle coordinate, p is the anisotropic pixel size, C_S is the spherical aberration, λ is the electron wavelength, f_N is the spatial Nyquist frequency, z is the anisotropic defocus value, and x is the sampling radius coordinate. A similar formulation was provided before⁶⁸ for the special case of isotropic pixel size, and

was used to reduce the comparison between CTF and PS to a 1D problem. However, *Warp* performs the fitting in 2D and only uses the rescaling for visualization purposes. If the complex defocus model fits the data better, the recovery of additional high-resolution Thon rings can be observed in the 1D average.

2.2.6 Resolution estimation

To estimate the useful resolution range, a normalized cross-correlation value between the averaged 1D PS and the simulated 1D CTF curve is calculated within a sliding window. The window size at any given position scales to twice the width between the zero points of the closest CTF peak, but is not allowed to fall below 16 samples. The resolution limit is then reported as the frequency where the cross-correlation falls below 0.3 for the first time. Since the higher number of optimizable parameters allows for some overfitting, it is important that the useful resolution extends beyond the range used for fitting.

2.2.7 Contrast transfer function estimation in tilt series

The single micrograph CTF estimation procedure with planar sample geometry described in the previous section can be used for tilted 2D data collection. However, full tilt series pose additional challenges for CTF fitting. Mechanical stage instabilities and imperfect eucentric height setting necessitate additional exposures for tracking and focusing⁶⁹ to correct the stage position between individual tilt images. Thus, the defocus cannot be assumed to stay constant, or change smoothly over the course of a tilt series. Each tilt image requires its own defocus value, which can be challenging due to the small amount of signal available. Even at $120 \text{ e}^-/\text{\AA}^2$ for an entire series of 60 images, each tilt only has $2 \text{ e}^-/\text{\AA}^2$ to perform the same estimation as for a $40 \text{ e}^-/\text{\AA}^2$ 2D image, while striving to achieve comparable accuracy.

CTF estimation in tilt series has traditionally received less attention than its equivalent in 2D data, with the most recent publication⁷⁰ predating the introduction of direct electron detectors and phase plates. As the resolution obtainable through sub-tomogram averaging has come close to parity with 2D data⁷¹ since then, simplifying assumptions such as the neglect of astigmatism⁷² or the assumed flatness of the sample can limit the

resolution. Combined with the lack of integration of dedicated tilt series CTF estimation tools into common sub-tomogram averaging pipelines³⁸, this has created a situation where state-of-the-art studies^{71, 73} employ tools designed for 2D data such as CTFFIND²⁷.

To improve the fit accuracy, the individual tilt image fits must be subjected to a common set of constraints. As the imaged sample content does not change significantly throughout the tilt series, 1D background and envelope can be derived from the average 1D spectrum of all tilt images. The relative tilt angles and the tilt axis orientation are known to a higher precision than could be derived from fitting a planar geometry *de novo*, and are kept constant throughout the optimization as suggested previously⁷². However, the absolute inclination of the sample plane is unknown. This is especially critical in some of the typical applications of tomography, like the imaging of lamellae prepared through FIB milling because a lamella can be tilted by over 20° relative to the grid. This additional inclination remains constant throughout the tilt series, and is made a single optimizable parameter for all tilt images. Astigmatism and, optionally, phase shift can be kept constant throughout 2D image exposures, but can benefit from a temporally resolved model in a tilt series where the overall exposure is fractionated over a much longer time, e. g. 20—30 min. *Warp* uses 3 control points in the temporal dimension to model these parameters.

With these considerations, the full estimation process is as follows. 2D patches are extracted from all aligned tilt movie averages, as described in the micrograph CTF fitting procedure, and treated in parallel in all subsequent calculations. To provide a better initialization for the per-tilt defocus searches, an estimate for the average defocus of the entire series is obtained by preparing 1D spectra from all patches, and comparing them to simulated CTF curves for the defocus values at the respective positions and tilts, taking into account the fixed relative tilt information and the currently tested average defocus (and phase shift, optionally). This search is performed exhaustively over a range of values specified by the user. The result is used as the starting point of a more complex optimization. Defocus values for all individual tilts, 3 astigmatism magnitude/angle pairs,

3 optional phase shift values, and the two global inclination angles (i. e. the plane normal) are optimized using the L-BFGS algorithm with the derivatives obtained numerically as described in the micrograph CTF fitting section. Upon convergence, the 1D spectra of all patches are rescaled to a common defocus value. This is especially useful for validation in tilt series since the individual images will have very noisy spectra. If the useful resolution range does not extend sufficiently beyond the fitting range, the latter is automatically decreased and the optimization repeated.

In our experience, the direction of the tilt axis is often miscalibrated. Correct handedness in structures obtained from sub-tomogram averaging does not guarantee the tilt angle sign is not flipped. In Warp, a positive rotation around the positive Y image axis is defined to result in an increased underfocus at positions to the positive X side of the tilt axis, i. e. those parts of the sample come physically closer to the electron beam source. The CTF fitting procedure in Warp can detect such mistakes by optionally repeating the optimization with the tilt angles flipped, and notifying the user if the “wrong” set of angles provides a better fit. Such a test can be used to re-calibrate the acquisition software for future data collection.

2.2.8 Considerations for tomogram reconstruction

Whereas the process of 3D map reconstruction from 2D images of single particles is well established today, full-tomogram reconstruction breaks some of the simplifying assumptions so they must be handled explicitly to obtain better results. In the 2D case, the CTF can be assumed to be the same for all parts of a single particle image, although corrections for a wider range of defocus values in images of large objects have been proposed⁷⁴. In a tomographic tilt series, the highest tilt image can show a defocus spread of 1 μm or more. Accounting for such variations in local defocus is necessary for reaching high resolution⁶⁵. Furthermore, each region in the tomographic volume is reconstructed from images with different CTFs, and the zeros and peaks of those CTFs will not overlap in Fourier space. CTF-based weighting of individual projections is commonplace for 2D data^{37, 75}, but the algorithms used in tomographic reconstruction do not go beyond CTF phase flipping,

giving all spectral components equal weight^{65, 76}. This gives spectral components with pure noise (CTF = 0) the same weight as the best available signal ($|CTF| = 1$) if they overlap. Performing CTF-, dose- and tilt-based weighting later in sub-tomogram averaging has been shown to be beneficial³⁸, but it has an even more significant effect when applied at the level of initial tomogram reconstruction. Anisotropic magnification has been described in the past²² and is routinely corrected in 2D data. In tomography, the real-space distortion is even more pronounced than in single particle reconstructions because the distances affected by the distortion are more than 1 μm , i. e. the extent of the entire tomogram, leading to positional errors on the order of nanometers in scenarios where the anisotropy does not coincide with the tilt axis.

2.2.9 Tomogram reconstruction

Warp takes the local defocus and sample distortion, as well as magnification anisotropy into account when reconstructing full or partial tomographic volumes. For a partial reconstruction at any position in the volume, the original 2D images are sampled at the following positions:

$$\mathbf{s} = \mathbf{R}_{Euler}(0, \alpha_{Tilt}, \psi_{Tilt}) \cdot \mathbf{p} + \mathbf{o}_{Tilt} ,$$

where \mathbf{R}_{Euler} is the rotation matrix for 3 Euler angles following the Xmipp convention⁷⁷, α is the stage tilt angle, ψ is the in-plane angle of the tilt axis, \mathbf{p} is the particle position within the tomographic volume, and \mathbf{o} is the in-plane offset of the tilt axis. The coordinates are centered within the volume and images. The CTF for each 2D image is calculated using a defocus of:

$$z = z_{Tilt} + \frac{\mathbf{s} \cdot \mathbf{n}_{Tilt}}{\mathbf{n}_{Tilt,z}} ,$$

where z_{Tilt} is the average defocus estimated for the tilt image, \mathbf{n}_{Tilt} is the sample plane normal, $\mathbf{n}_{Tilt,z}$ is the z component of the normal, and $*$ denotes the scalar product between two vectors. The reconstruction is performed in Fourier space using a gridding algorithm³⁷, with the data weighted by the respective CTF, and the dose- and tilt-dependent heuristic from RELION³⁸, but without the final deconvolution step (i. e. the weights

are inserted as $|CTF|$, not as CTF^2). To obtain a full tomogram, *Warp* reconstructs a uniform grid of small, cubical volumes with an overlap of 50%, and inserts the central 50% into the overall volume to account for artifacts associated with Fourier space reconstruction at the borders of the local volumes. This ensures the corrections can be applied with local precision and remain reasonably continuous between adjacent sub-volumes.

2.2.10 Export of corrected data

Whereas the *Warp* model for a movie or tilt series describes the non-linear deformation of the entire particle ensemble and its environment, it is unclear whether this deformation gradient stays continuous throughout a single particle, i. e. if the protein structure is subject to the same compression and shearing as the ice around it. Many recent high-resolution maps were reconstructed using particles extracted from dose-weighted averages produced by MotionCor2²³. The tool assumes the deformation gradient to be continuous in all parts of the image, and will thus deform images of particles and ice in the same way. This will be beneficial if the underlying physical model is indeed continuous. However, it also distorts the CTF locally without passing any knowledge of the distortion to downstream processing tools. In case of a strong local change in the motion direction, this will result in an artifact similar to lens astigmatism.

Warp assumes a continuous deformation field when exporting dose-weighted averages of whole 2D movies, i. e. each pixel will be shifted according to the grid interpolants at that exact position. This has the benefit of uniformly sharper images for visual inspection and particle picking. For particle and sub-tomogram extraction, however, the entire particle image will be shifted uniformly according to the grid interpolants at the particle's center. This keeps the CTF true to its fitted analytical description, but makes the assumption that the protein is more rigid than the surrounding ice and thus deforms less due to BIM. For whole-tomogram reconstruction, a hybrid approach is pursued: the local volumes are produced using the same procedure as sub-tomogram extraction, but the combined volume is largely continuous depending on how small the local volumes were.

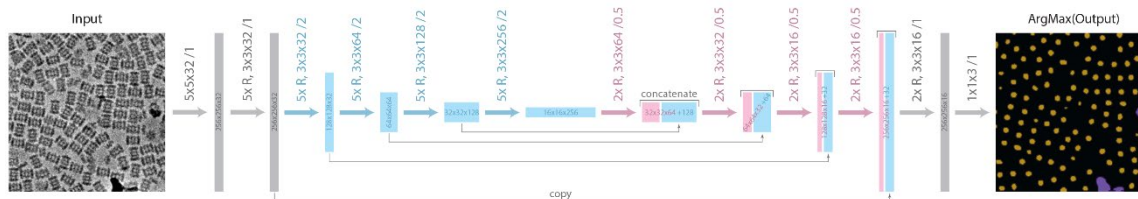


Figure 2.12 | Neural network architecture of BoxNet.

Rectangles depict the intermediate tensor dimensions. Their width and height are proportional to the number of channels and the spatial extent, respectively. Thick arrows represent convolution operations. Their format is encoded as “(Kx R), LxMxN /O”, where K is the number of consecutive ResNet blocks, or absent in case of a single convolution operation; L and M are the dimensions of the convolution kernel; N is the number of kernels, resulting in N channels in the output; O is the stride length (1 = no change, 2 = downsampling by factor of 2, 0.5 = upsampling by factor of 2 through transposed convolution). The stride parameter is only applied to the first convolution in a chain of ResNet blocks, whereas all subsequent convolutions use stride = 1. The contractive part of the network is colored in cyan, the expanding part in magenta. The final image shows the result of applying a per-pixel ArgMax operator to the result of the last convolution to obtain the spatial distribution of the 3 labels the model is trained to predict: background (black), particle (yellow), artifact (purple).

Dose weighting in *Warp* adds a B-factor of -4 \AA^2 per $1 \text{ e}/\text{\AA}^2$ of dose, similar to a heuristic published previously⁷. While a different heuristic is used in MotionCorr²³ and Unblur⁷, the accuracy of both approaches is of decreased significance as data-driven re-weighting is likely to be performed using an approach like the “particle polishing” in RELION 3.0.

2.2.11 Particle picking with a residual neural network

In the past years, the recipe for improving the performance of deep learning algorithms has been “deeper networks, more training data”. Outside of cryo-EM, deep ResNet architectures have been demonstrated to enable the training of very deep models by essentially solving the vanishing gradient problem⁵¹. At the same time, the EMPIAR raw data repository⁵⁸ has accumulated a diverse collection of 2D cryo-EM data sets that can be leveraged for training. *Warp* employs a model with 35 ResNet blocks and 2 conventional convolution layers (**Figure 2.12**) to segment a micrograph into 3 classes: background, particle, and high-contrast artifact (e. g. ethane drops). The input window has a constant size of 256^2 px. After initial convolution with $32 \text{ } 5 \times 5$ kernels the data are processed by a

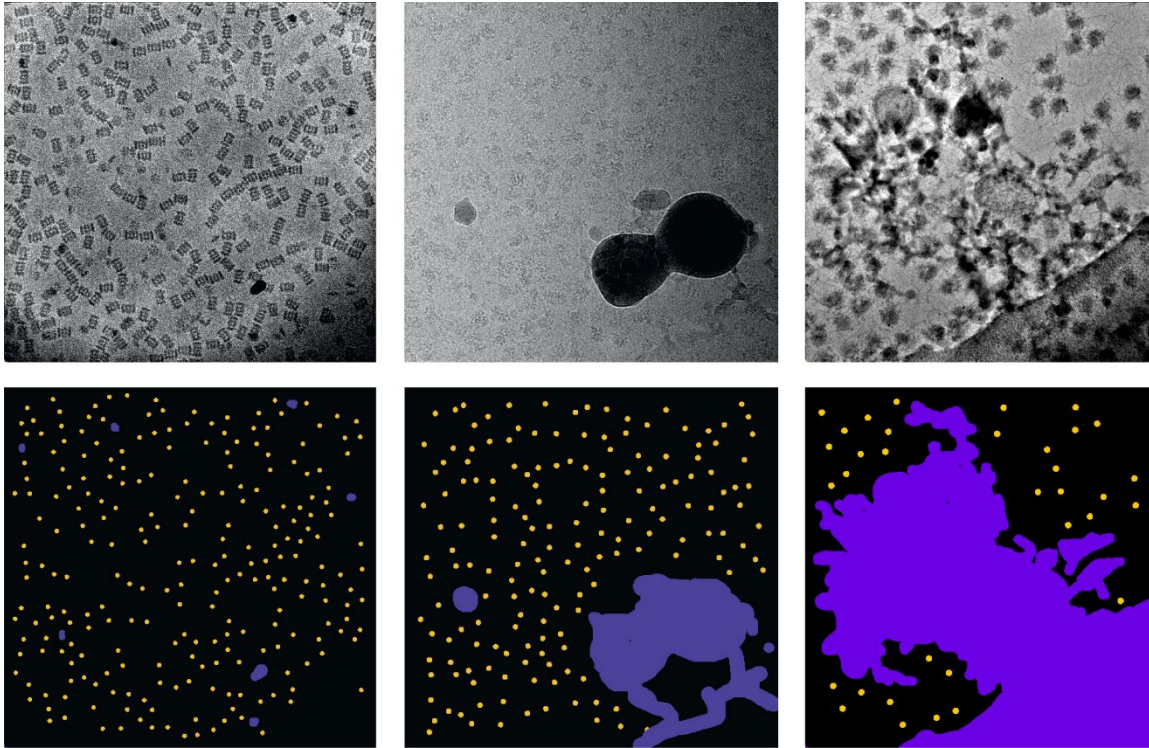


Figure 2.13 | Examples of data used to train BoxNet.

Examples of micrographs presented to BoxNet as input (top row), and the per-pixel labels used as the desired output during training (bottom row). The pixel classes predicted by BoxNet are background (black), particles (yellow), and artifacts (purple).

sequence of 5 groups of each 5 ResNet blocks. At the beginning of each but the first group, the data are down-sampled by a factor of 2, while the number of channels is doubled. This enables the recognition of an increasing number of large, higher-order features. After reaching a spatial extent of 16x16 in the contractive part of the network, the data are processed by the expanding part: a sequence of 4 groups of each 2 ResNet blocks. At the beginning of each group, the data are up-sampled by a factor of 2. Each group's output is concatenated with the output of its mirror counterpart from the contractive part in a U-Net-like fashion⁵⁰. This combines the global context and higher-order features obtained in the contractive part with the higher spatial resolution of the previous layers. After reaching the original extent of 256x256 in the expanding part, the data are processed by the final 2 ResNet blocks and projected onto 3 channels by convolution with 3 1x1 kernels. A pixelwise argmax operation finally retrieves the most likely label at each

location in the original image. The model graph and variables are initialized and saved using TensorFlow's Python API, while all subsequent training and inference are performed through its C API, wrapped in C# classes.

To segment a micrograph, it is down-scaled to the standard 8 Å/px resolution and divided into 256^2 px tiles with 64 px overlap. Each tile is extracted, normalized, and presented to the network. All tiles' softmax and argmax outputs are combined and subjected to a user-defined threshold for the softmax value to remove uncertain picks. Connected components of pixels labeled as "particle" are extracted, and their centroids are used as particle positions. In cases where two particles overlap according to a user-defined diameter, the particle with the bigger connected component is kept. To obtain a mask from the "artifact" label, a threshold of 0.1 is applied to the softmax values, and connected components with less than 20 pixels are removed. The remaining pixels are saved as a binary mask. The final list of particle positions only contains those with a user-defined minimum distance to the masked regions.

2.2.12 Initial training of BoxNet

11 EMPIAR and 14 in-house data sets (**Figure 2.13, Table 2.1**) each contributed 20–50 micrographs to the training set. Additionally, synthetic data were prepared from 21 PDB models (**Table 2.1**) using a modified version of the InSilicoTEM⁷⁸ package, contributing ca. 1600 particles per species. The simulated data contained only one species per micrograph, although more heterogeneous examples might be added in the future. The training set was split 9/1 for training/validation, and trained with the momentum optimizer in TensorFlow 1.5 using a learning rate gradually decreasing from 1E-2 to 1E-5. The normalized data were augmented in each training epoch by extracting the 256^2 px window at random positions, and applying random rotation, flipping, shearing, and Gaussian noise with a random standard deviation between 0.0 and 0.6. This augmentation was observed to have an excellent regularizing effect, as the final training and validation scores were virtually identical. The training was performed for 800 epochs, using a batch size of 1.

Access code	Sample	Synthetic	Carbon support	Phase plate
EMPIAR-10017	beta-galactosidase			
EMPIAR-10077	80S ribosome	✓		
EMPIAR-10078	20S proteasome		✓	
EMPIAR-10081	HCN1 channel			
EMPIAR-10084	Haemoglobin		✓	
EMPIAR-10089	TcdA1 in prepore state			
EMPIAR-10097	Influenza Hemagglutinin			
EMPIAR-10122	Apoferitin		✓	
EMPIAR-10153	80S ribosome	✓	✓	
EMPIAR-10156	80S ribosome	✓		
EMPIAR-10200	Apoferitin			
	RNA Polymerase II complex			
	RNA Polymerase II complex	✓		
	RNA Polymerase II complex			
	Viral polymerase			
	Nucleosome complex			
	Chromatin-related	✓		
	Chromatin-related			
	Transcription-related complex			
	Transcription-related complex			
	Transcription-related complex	✓		
	Chromatin-related			
	Chromatin-related	✓		
	RNA Polymerase II complex			
	Nucleosome			
PDB-1sa0	Tubulin-Colchicine	✓		
PDB-2gtl	Lumbricus Erythrocrutorin	✓		
PDB-2wri	70S ribosome	✓		
PDB-3j9i	20S proteasome	✓		
PDB-4hhb	Haemoglobin	✓		
PDB-4zor	S37P MS2 viral capsid	✓		
PDB-5foj	Grapevine Fanleaf virus	✓		
PDB-5mmi	Chloroplast ribosome, large subunit	✓		
PDB-5ngm	70S ribosome	✓		
PDB-5vy5	Aldolase	✓		
PDB-5w3l	Rhinovirus B14	✓		
PDB-5w3s	TRPML3 channel	✓		
PDB-5xnl	Stacked PSII-LHCII supercomplex	✓		
PDB-5xwy	LbuCas13a-crRNA complex	✓		
PDB-5y6p	Phycobilisome	✓		
PDB-6az1	80S ribosome, small subunit	✓		
PDB-6b7n	Coronavirus spike protein	✓		
PDB-6b44	CRISPR Csy surveillance complex	✓		
PDB-6bco	TRPM4 channel	✓		
PDB-6bcx	mTORC1	✓		
PDB-6bhu	MRP1	✓		

Table 2.1 | Experimental and synthetic data used to train the general BoxNet model.

2.2.13 Retraining of BoxNet

User-supplied positions of positive examples and, optionally, areas of increased and decreased certainty in the micrographs are automatically converted to training data. If requested, the training set is diluted with data from the latest version of the centrally curated set (in the following referred to as 'old data') in a 1:1 ratio to prevent possible overfitting of the new data. The retraining regime is identical to initial training, but lasts only 100 epochs. During the retraining, 4 metrics are calculated continuously for every batch: the old network's accuracy for old and new data, and the retrained network's accuracy for old and new data. Ideally, the retrained network's accuracy for new data will improve to approach or even surpass the old network's accuracy for old data by the end of the retraining process, whereas the accuracy for old data will stay constant.

2.2.14 Template matching in micrographs and tomograms

2D micrographs are subdivided in tiles with an overlap matching twice the template particle diameter. For each square, 2D projections of the template are prepared at user-defined angular intervals, convolved by the square's CTF, and normalized to mean = 0, standard deviation = 1 in real space. The square's FT is multiplied by the conjugate of the projection's FT, and an IFT yields the cross-correlation scores for all positions within the square. These scores are normalized by the local standard deviation within the square. The scores are compared for all template orientations, and the best one is stored for each pixel within the square. Finally, the result is cropped to exclude a border matching the template particle diameter, and combined with the results from other squares to obtain the correlation scores for the entire micrograph. A local peak search is performed using the template particle diameter as the minimum distance, and all peak positions are stored for further processing. Template matching in tomographic volumes follows the same concept. Instead of square tiles, local cubes are cross-correlated with the template convolved by the local 3D CTF. Optionally, a spectrum whitening of the target micrograph/tomogram can be performed as previously described⁷⁹. This has the benefit of

equalizing the spectral noise amplitudes for all spatial frequencies, effectively giving more weight to the higher frequencies and sharpening the correlation peaks.

2.2.15 Deconvolution

In the absence of a phase plate, the CTF will be dominated by its sine component, i. e. have very little contrast in the lowest spatial frequencies. This creates a high-pass filter effect in the raw data and, due to increasingly noisier higher frequency components, makes it hard to assess the image content visually. On the other hand, a phase plate creating the desired phase shift of $\pi/2$ will apply a low-pass filter in defocused images, rendering them blurry. Both scenarios do not affect subsequent alignment and averaging procedures significantly, and the filters will be reversed in the final reconstruction by dividing its 3D FT by the weighted average of all contributing CTFs. This becomes possible because the spectral signal-to-noise ratio (SSNR) is sufficiently high after averaging enough particles with different CTFs. However, even in single images the lowest frequency components often contain enough signal so that boosting them by inverting the CTF will increase the visible low-frequency contrast while maintaining acceptable noise levels. This provides conventional images with a better definition of object boundaries, making their manual selection easier. In defocused phase plate images, this improves sharpness.

To construct a Wiener-like filter, *Warp* makes ad hoc assumptions about the SSNR that can be adjusted by the user. The SSNR is assumed to be a combination of an exponential decay curve and a raised cosine high-pass filter:

$$SSNR_f = e^{-f \cdot 100F} \cdot 10^{3S} \cdot H_f ,$$

where f denotes the spatial frequency, H is an optional high-pass filter, F is the custom fall-off parameter, and S is the custom strength parameter. The factors for F and S are empirically tuned so that the default values of 1 produce good results for typical direct electron detector data, although adjustments might be required in some cases. The SSNR is then used in a Wiener-like filter:

$$I'_f = \frac{I_f \cdot CTF_f}{CTF_f^2 + SSNR_f^{-1}},$$

where I is the FT of the image, and CTF is the 2D contrast transfer function. The shape of the SSNR curve prevents the lowest frequency components from being boosted too much, giving rise to a noisy sample background, and acts as a low-pass filter at the same time to suppress the noisy high frequency components. An example of such a filter and its effect on a 2D micrograph are shown in **Fig. S2b**. In practice, a higher electron dose helps to obtain good low-frequency contrast in conventional images. The commonly used dose of 30–40 e[−]/Å² works well for holey grids with thin ice, while more might be required in the presence of carbon support or thick ice. The deconvolution works especially well in tomograms, where the overall dose often surpasses 100 e[−]/Å².

2.2.16 Denoising improves particle visibility

Even after deconvolution, micrographs will often display a high amount of noise that makes their visual inspection difficult. Deep convolutional neural network-based denoisers have been shown to perform better than hand-crafted statistical models^{80, 81}. [REFs]. However, their training has traditionally required pairs of noisy and noise-free observations of the same signal. This prevented denoiser training on cryo-EM data, where a noise-free ground truth is not available due to quickly increasing radiation damage in sensitive samples. Recently, the Noise2Noise principle⁸² has been proposed to circumvent this limitation while achieving comparable performance. Pairs of independently noisy observations of the same signal are used in training as input and output. In the absence of correlation between the noise patterns, the expectation value is the underlying noise-free signal, which the model learns to predict. Pairs of independently noisy data are readily available in cryo-EM. Since every micrograph's signal is fractionated in a multitude of frames, computing the aligned averages of all odd and even frames renders a pair of almost identical observations with independent patterns of camera shot noise. By using odd and even frames instead of the first and second halves of a movie, the effect of radiation damage is very similar in both averages.

Warp adopts the published Noise2Noise U-net-like network architecture with one change; a batch normalization layer is added before each leaky ReLU layer. The denoising is performed on overlapping 256^2 px tiles, and the central 128^2 px windows are combined to produce the final image shown in Fig. S2c. *Warp* generates training data from processed movies automatically. These images are also pre-deconvolved. While deconvolution and denoising are independent processes, their combination provides the most natural-looking result for the human observer. Once enough training examples have been generated (up to 128 are supported by default), the user can request to retrain the model, which will take a few minutes. A default pre-trained model is provided with *Warp* that covers a narrow range of conventional and VPP data. However, the best denoising performance is achieved by retraining the model for every new data set. Like deconvolution, denoising can be applied on-the-fly to any micrograph displayed in *Warp*, using the default or retrained model. While denoising provides a great visual aid for human interpretation of raw data, the process removes all signal that cannot be reliably distinguished from noise in a single observation. Thus, denoised particle images do not render themselves to averaging as a means of increasing the SNR at higher resolution.

2.2.17 Benchmarking for 2D data

For the influenza hemagglutinin trimer benchmark, raw movie data and pre-extracted particles from EMPIAR-10097 were downloaded. The movies were processed with the full *Warp* pipeline using the following settings: motion correction with a temporal resolution of 40 for the global motion, and 5x5 spatial resolution for the local motion, using the 0.03–0.25 Nyquist range and a B-factor of -400 \AA^2 ; CTF estimation with 6x6 spatial resolution, using the 0.1–0.35 Nyquist range; particle picking with a BoxNet model retrained on particles from 3 micrographs, using the default 0.95 threshold. Quality filters were applied in *Warp* as follows: defocus between 0.3 and $5.0 \text{ }\mu\text{m}$, resolution better than 8 \AA , intra-frame motion of at most 1.5 \AA , particle count above 120. Particles were extracted from the micrographs meeting these filters and subjected to processing in cryoSPARC: no 2D classification was performed; *ab initio* refinement was performed with 6

classes and no symmetry; the 6 classes were then refined heterogeneously, with no symmetry imposed; the only class showing the expected Hemagglutinin structure was refined with C3 symmetry. The original particle set from EMPIAR-10097 was subjected to 3 different processing strategies. First, the full set was refined in cryoSPARC with C3 symmetry using the original CTF estimates. Second, the full set was subjected to the same classification and refinement as the particles from Warp, using the original CTF estimates. Third, particles from the Hemagglutinin class obtained in the second processing branch were updated with local CTF estimates from Warp, and refined again with C3 symmetry. Resolution estimates were obtained for all maps using the respective masks automatically generated by cryoSPARC.

For the β -galactosidase benchmarking studies, raw data from EMPIAR-10061 were downloaded. The movies were processed with the full *Warp* pipeline using the following settings: motion correction with a temporal resolution of 38 for the global motion, and a 5x5 spatial resolution for the local motion, using the 0.03–0.60 Nyquist range and a B-factor of -160 \AA^2 ; CTF estimation with 5x5 spatial resolution, using the 0.08–0.60 Nyquist range; particle picking with a BoxNet model retrained on particles from 5 micrographs, using a threshold of 0.30. No quality filters were used as the data already represent a high-quality subset curated for the initial publication. Picked and extracted particles were subjected to 2D and 3D classification with C1 symmetry in RELION 2.1 to remove incomplete particles. The remaining particles were refined with D2 symmetry. The final half-maps were then used to refine beam tilt and per-particle defocus in RELION 3.0. Global motion tracks for all movies were exported from *Warp* to RELION 3.0 to perform Bayesian particle polishing.

To assess the frame alignment accuracy in *Warp* independently of downstream map refinement, β -galactosidase movies were aligned in *Warp* as described above, and using the default settings in MotionCor2. CTF fitting was performed with 5x5 spatial resolution, using the 0.08–50 Nyquist range. Frequency-dependent fit quality was calculated as

described in the 'Resolution estimation' section, and all resulting curves averaged. The resolution was then estimated at a cut-off value of 0.3.

2.2.18 Benchmarking for tilt series data

For the template matching benchmark, pre-aligned tilt series from EMPIAR-10045 were downloaded. The defocus was estimated, and full tomograms were reconstructed with a pixel size of 10 Å in *Warp*. The 80S ribosome map derived from these data in the original publication⁶⁴ and deposited under EMD-3228 was used as the template. Template matching was performed on the 10 Å/px tomograms with an angular sampling of 7.5 °, using a local 3D CTF. The same steps were performed using a binary missing wedge mask instead of the 3D CTF. The picked positions were screened manually to determine the false positive rate. Background statistics were calculated for the correlation volume trimmed to remove the vacuum region, and excluding 48 px windows around the peaks.

To benchmark CTF estimation and sub-tomogram export on EMPIAR-10045 data, the particles previously picked through template matching were exported together with their 3D CTF volumes at a pixel size of 5 Å. The sub-tomograms were then subjected to 3D refinement in RELION 3.0 without prior classification.

To benchmark CTF estimation and sub-tomogram export on HIV-1 particles, raw data from EMPIAR-10164 were downloaded. A subset of 5 tilt series previously used by the authors of NovaCTF⁶⁵ was selected. Movies were aligned in *Warp* using only global alignment with a temporal resolution of 5. Gold beads were picked manually and used to align the tilt series in IMOD⁸³. Full tomograms were reconstructed with a pixel size of 5 Å in *Warp*. Template matching was performed with the EMD-4015 map with an angular sampling of 7.5 ° and C6 symmetry. A custom script was used to remove particles not fitting into a regular hexagonal grid as described previously⁷¹. The particles were exported together with their 3D CTF volumes at a pixel size of 1.35 Å. The sub-tomograms were then subjected to 3D refinement in RELION 3.0 without prior classification.

3. Multi-particle refinement in *M*

The work presented in this chapter was published in:

D. Tegunov, L. Xue, C. Dienemann, P. Cramer, J. Mahamid. Multi-particle cryo-EM refinement with *M* visualizes ribosome-antibiotic complex at 3.7 Å inside cells. bioRxiv 2020.06.05.136341, Nature Methods ('accepted in principle').

3.1 Results

3.1.1 Overall design

M was designed to form the last part of a largely automated cryo-EM data pre-processing and map refinement pipeline – preceded by Warp⁸⁴ and RELION³⁷, or compatible tools (**Figure 3.1**). Warp performs initial, reference-free motion correction and CTF estimation on frame series or tilt movies during data acquisition. For tilt series pre-processing, Warp, starting with version 1.1.0, automatically calls routines from IMOD⁸³ to perform the initial tilt series alignment, estimates per-tilt CTF using the tilt angles as constraints, and reconstructs the tomographic volumes at a large pixel size for visual analysis and particle picking. Warp then picks the particles using a convolutional neural network-based (CNN) approach for frame series, or template matching for micrographs or tomograms, and exports them as images or reconstructed volumes depending on the data type. In case of tilt series, 3D CTF volumes containing the missing wedge and tilt-dependent weighting information are generated for each particle³⁸. The particle poses and classes are then determined in RELION using a multitude of strategies available there⁸⁵. All classes and their respective refinement results are finally imported into *M* to perform a more accurate, reference-based, multi-particle frame or tilt series refinement and obtain the final high-resolution maps. Optionally, the refined parameters can be used to re-export more accurately aligned particles for further classification in RELION or compatible software. The new alignments can be applied to generate tomograms at higher resolutions to be used for further particle picking.

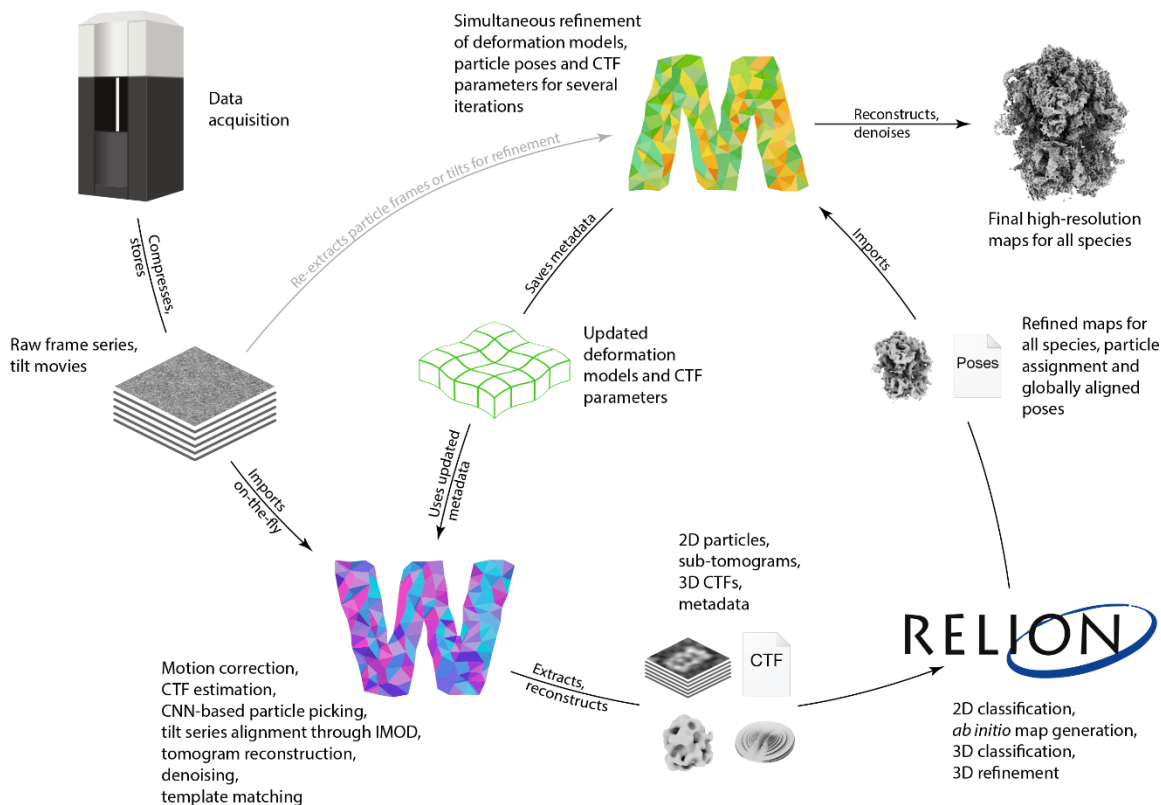


Figure 3.1 | The Warp-RELION-M pipeline for frame and tilt series cryo-EM data refinement.

Electron microscopy data are pre-processed on-the-fly in Warp, which then exports particles as images or sub-tomograms. Particles are imported in RELION, where they can be subjected to a multitude of processing strategies, resulting in 3D reference maps, global particle pose alignments, and class assignments. The particle population encompassing all classes is then imported in M, where reference-based frame or tilt image alignments are performed simultaneously with further refinement of particle poses and CTF parameters to improve resolution. Finally, M produces high-resolution reconstructions that can be used for model building. Alternatively, the improved alignments can be used in Warp to re-export particles for further, more accurate classification in RELION.

M provides a graphical user interface (GUI) that allows the user to create, import, export and manage data. Projects are organized as “populations”, which contain “data sources” and “species”. A data source is a set of frame or tilt series, that stem ideally from the same sample grid and acquisition session. A species is any distinct type of macromolecule, or its compositional and conformational sub-state. Each version of a data source or species is tracked using a cryptographic hash of its current state, the preceding version, and

the processing parameters connecting them. The entire refinement evolution can be tracked as a directed graph, parts of which can be stored in different locations while remaining uniquely connected through the hashes. Thus, multiple users can process different collections of data sources and species independently and merge them later. This is especially useful for processing data sets of complex cellular environments, where each user will typically process only a small fraction of all species contained and everyone would benefit from pooling data and results together.

3.1.2 Multi-particle system modeling

M considers the entire field of view of a frame or tilt series as a physically connected multi-particle system (**Figure 3.2a**). The particles can belong to different refined species, which can be of varying size, symmetry, and resolution. As parts of the same system, the particles are subject to the same global transformations such as the translation and rotation of the microscope stage, as well as locally similar transformations caused by BIM that result in apparent translation and rotation of particles. *M* performs a reference-based registration of these transformations (**Figure 3.2b**), and reverses them when back-projecting individual particle frame or tilt images to obtain more accurate reconstructions.

In frame series, all transformations occur in the same image reference frame. Their combined effects are parametrized as a pyramid of 3D cubic spline grids (**Figure 3.3**). This pyramid results from a combination of grids where the top grid has low spatial and high (per-frame) temporal resolution, and 1–2 subsequent grids have double the spatial and 1/4 the temporal resolution of the previous grid. This model is similar to the one used in Warp but fits more parameters due to the higher accuracy of reference-based registration. The user can set the spatial resolution of the top grid to adjust the model’s resolution to the available signal. In addition to image-space warping, *M* can fit doming-like motion that is known to occur at the beginning of an exposure⁹ (**Figure 3.2b**). This is implemented as parameter grids for defocus and orientation offsets with 3x3 spatial and per-frame temporal resolution.

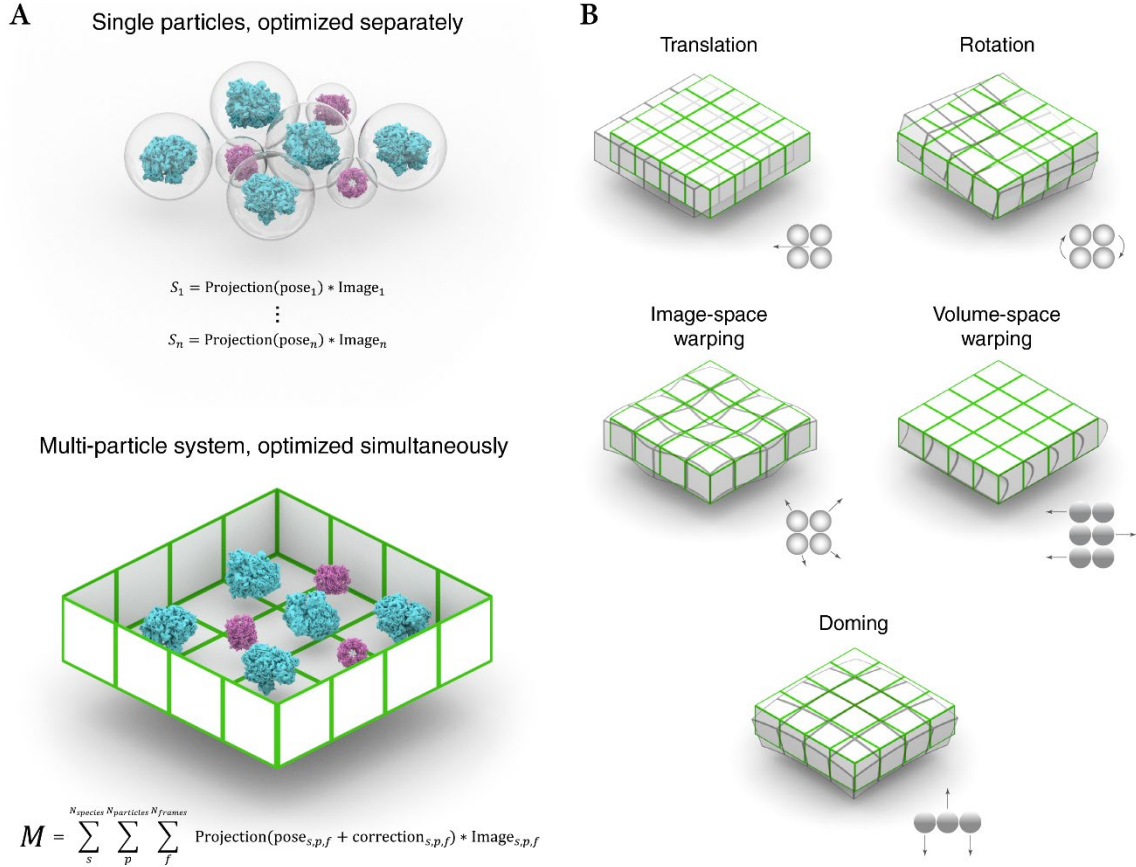


Figure 3.2 | Multi-particle system modeling and optimization.

M employs a reference-based multi-particle optimization to model sample deformation and improve map resolution.

a) Particles are typically treated as isolated entities in SPA. Each particle has its own cost function based on the similarity between a simulated reference projection and the experimental particle image, which is optimized independently. However, particles in a real sample are physically connected and experience locally similar effects during exposure. Each imaged location is modeled as a multi-particle system. Its state model is fitted using a single cost function, which compares simulated reference projections to all experimental particle frame or tilt images. The particle poses in each frame or tilt are additionally affected by the modeled deformation of the multi-particle system, which is optimized together with the per-particle pose alignments.

b) The multi-particle system deformation model incorporates several modes: Global movement and rotation to account for inaccuracies in stage movement between frames and stage rotation between tilts; image-space warping to model local non-linear deformation in the 2D reference frame of a frame or tilt image; volume-space warping to model the movement of overlapping particles perpendicular to the projection axis (tilt series only); doming to account for the hypothesized bending of a thin sample along the projection axis (frame series only).

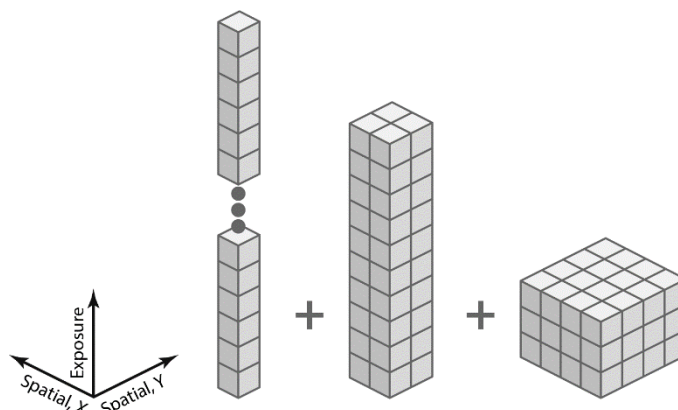


Figure 3.3 | Example of a parameter grid pyramid that models in-plane motion in a frame series.

Several grids are combined to model the in-plane motion occurring in a frame series with 40 frames as a function of position and dose. Each cubical cell represents a sampling point. Starting with a grid with full temporal (exposure) and no spatial resolution to model fast, global motion (left, 1x1x40, shown truncated), temporal resolution is reduced by a factor of 4 and spatial resolution is doubled to model slower, local motion (center, 2x2x10; right 4x4x3). The spatial resolution of the first grid can be set higher if there is enough particle signal to fit.

For tilt series data, *M* distinguishes image-space and volume-space effects because the tilt images show the volume from different angles. Image-space transformations are parametrized as a 3D cubic spline grid with per-tilt temporal, and a spatial resolution set by the user depending on the available signal. Additionally, parameters of a coarse 3D cubic spline grid can be fitted for every tilt movie to account for the significant exposure and deformation captured in each of them. Volume-space transformations, such as the shearing of a thick sample, are modeled as a 4D parameter grid with quadrilinear interpolation, with the accumulated exposure as the temporal dimension. Because *M* does not average particle frames or tilts in intermediate steps, per-particle translation and rotation trajectories can be fitted to model the most local transformations. The temporal resolution of the trajectories can be set for each species depending on its size and thus the signal available per particle.

Using frame series data collected on an apoferritin sample (AF-f, see Methods), we show the benefit of considering the particles of multiple species in refinement. To this end, we

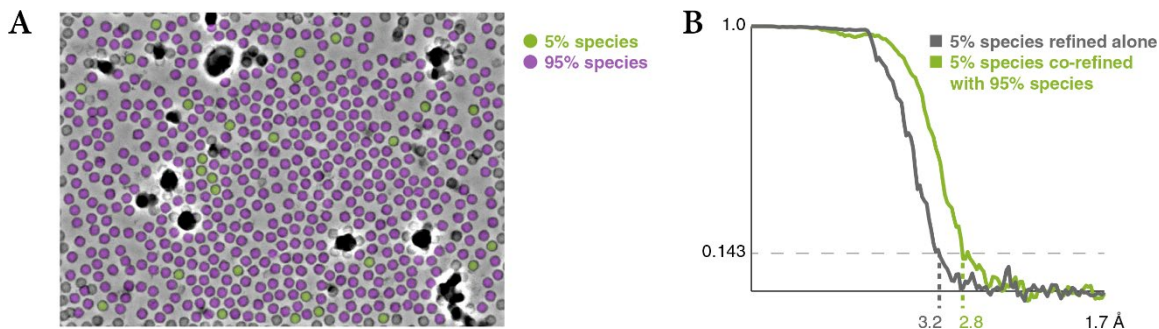


Figure 3.4 | Benefits of considering more particles per micrograph through multi-species refinement.

Apoferitin frame series were refined using a small 5% sub-population of the particles alone, and together with another 95% sub-population that improved the accuracy of the multi-particle system hyperparameters, but did not contribute particles to the 5% half-maps.

a) Exemplary distribution of the 2 sub-populations within a frame series.

b) FSC curves between the half-maps of the 5% population in both scenarios, showing the benefit of multi-species refinement.

artificially split the apoferitin population in two species comprising 5% or 95% of the particles (**Figure 3.4a**, **Table S3.1**). No structural similarity between the two species was assumed during refinement. Refining the 5% species alone produced a 3.2 Å map, while adding the 95% species to the multi-particle system improved the map calculated from the 5% species to 2.8 Å (**Figure 3.4b**). This demonstrates that our multi-species refinement approach can improve the resolution for heterogeneous data sets.

3.1.3 Correction of electron-optical aberrations

In addition to a geometric deformation model, *M* fits CTF parameters and higher-order aberrations including beam tilt. For frame series, defocus is optimized per-particle, similar to cisTEM⁸⁶ and recent RELION versions³³. For tilt series, defocus is optimized per-tilt, similar to the capability offered in emClarity⁴³. For both types of data, astigmatism, anisotropic pixel size and higher-order aberrations are fitted and corrected per-series.

CTF correction at high defocus can introduce artifacts if the chosen particle box size is too small to retain high-resolution Thon rings, leading to their aliasing (**Figure 3.5a**) and limiting the resolution for many combinations of high defocus images and small particle

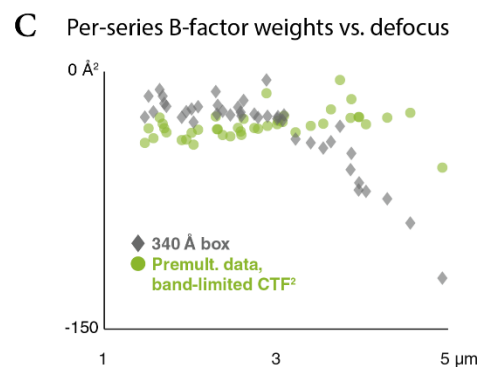
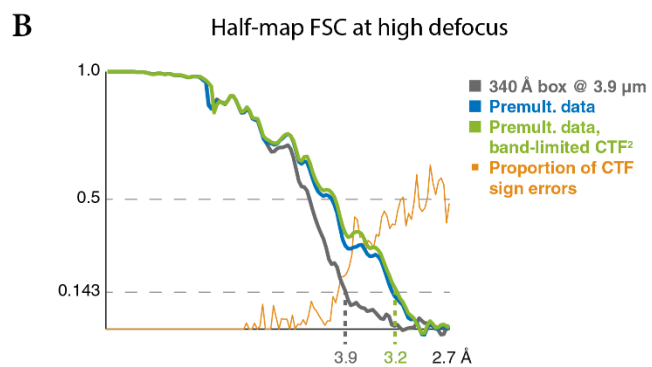
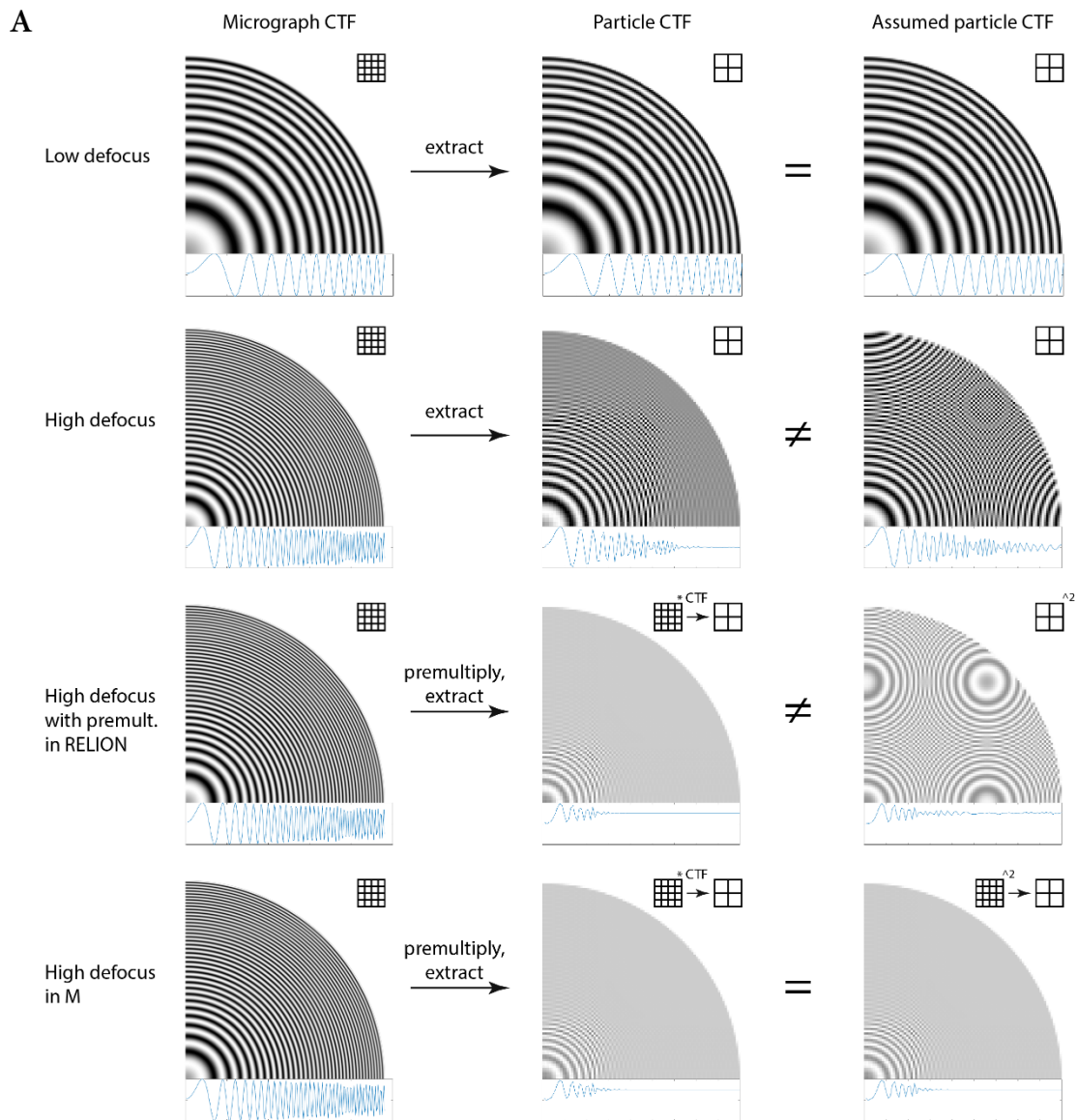


Figure 3.5 | CTF correction at low and high defocus.

High-resolution information is delocalized at high defocus. Choosing an insufficiently large particle box size results in loss of that information. In Fourier space, this results in CTF oscillations becoming too fast to be resolved at the sampling rate provided by the small box, averaging to 0. *M* chooses the box size automatically for each frame or tilt series' defocus, pre-multiplies the data and simulated CTF by the CTF to eliminate the oscillations and localize the signal, and then crops the data to the desired map size. This avoids the pitfall of losing map resolution due to an inappropriately chosen box size.

a) Visualization of the delocalization and aliasing effects in Fourier space as 2D and rotationally averaged 1D CTFs; grids depict sampling rate. At low defocus (row 1), all signal is localized within the box and no aliasing is seen in the simulated CTF used for the image formation model during refinement. At high defocus (row 2), high-resolution signal is delocalized outside the small particle box. Once the particle is extracted, the fast CTF oscillations are averaged to 0 and high-resolution information is lost. At the same time, the simulated CTF is filled with aliasing artifacts because it is not low-pass filtered in the same way. If the particle data are pre-multiplied by the CTF at a box size large enough to contain all signal and resolve all CTF oscillations (row 3), as can be done optionally in RELION, all particle signal is contained in the box after cropping it to a smaller size, and the CTF averages to 0.5. However, the simulated CTF² does not match this and contains aliasing artifacts. *M* applies the pre-multiplication to both particle data and simulated CTF in a larger box before cropping (row 4) to avoid the mismatch.

b) FSC between the half-maps reconstructed from HIV1 virus-like particles of a single high-defocus (3.9 μm) tilt series in an insufficiently large box. Using data extracted without pre-multiplication, as is currently common, limits the resolution to 3.9 \AA (grey). Pre-multiplying both particle data and CTF in a larger box, as automated in *M*, provides the best 3.2 \AA result (green). Pre-multiplying only particle data is only slightly worse here (blue), but would likely lead to noticeably worse results in RELION as the aliased CTF² would be used in the image formation during refinement. The FSC curves diverge as the proportion of CTF sign errors (orange) increases.

c) Relation between tilt series defocus and associated contribution of high-resolution information to the reconstruction. For the larger data set, not pre-multiplying the data results in a strong correlation, where high-defocus data is down-weighted to contribute less (grey). The correlation disappears when pre-multiplication is applied, so more tilt series contribute high-resolution information (green).

box sizes. *M* automates the selection of a sufficiently large box size at which the data are pre-multiplied by an aliasing-free CTF. The images are then cropped in real space. To match the underlying CTF of these images, correctly band-limited CTF² images are constructed in a similar way (**Figure 3.5a**). Both are then used for refinement and reconstruction.

We show the benefit of this approach by reconstructing a map from a previously refined high-defocus tilt series of HIV1 virus-like particles (EMPIAR-10164, **Table S3.1**). Using twice the particle diameter as the box size, the resolution is limited to 3.9 Å as the average sign error of the aliased CTF increases (**Figure 3.5b**). Pre-multiplying the data and CTF at an aliasing-free size and then cropping them improved the resolution to 3.2 Å using the same reconstruction box size. Only pre-multiplying the data but using an aliased analytical CTF² for the Wiener-like reconstruction filter did not decrease the nominal resolution in this case. However, for algorithms that would use such aliased models during refinement and classification as well, we expect these effects to be more noticeable. This approach improved the empirically estimated per-tilt series weighting factors (see Methods) for high-defocus data to the level of low-defocus data for the entire EMPIAR-10164 data set (**Figure 3.5c**).

3.1.4 Optimization procedure

M optimizes all selected hyperparameters describing geometric deformation (**Figure 3.2b**), electron-optical aberrations, and particle pose trajectories, simultaneously. Because exhaustive search over an ensemble of thousands of parameters would be impossible, *M* performs a local, gradient descent-type optimization using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm⁶⁶. The target function is the sum of normalized cross-correlations between all extracted particle images contained in a frame series or a tilt series, and reference projections at angles and shifts defined by the particles’ poses and additional corrections described by the hyperparameters (**Figure 3.2a**). To compute the derivatives for the variables efficiently, *M* precomputes sets of weights using a strategy similar to Warp’s⁸⁴ (see Methods). Derivatives for many of the parameters can then be computed as weighted sums of per-particle image derivatives, which in turn are calculated using GPU-accelerated routines. The optimization procedure considers the signal of all defined particle species simultaneously to maximize the particle density in each frame or tilt series, thus increasing the hyperparameter fitting accuracy for heterogeneous data sets.

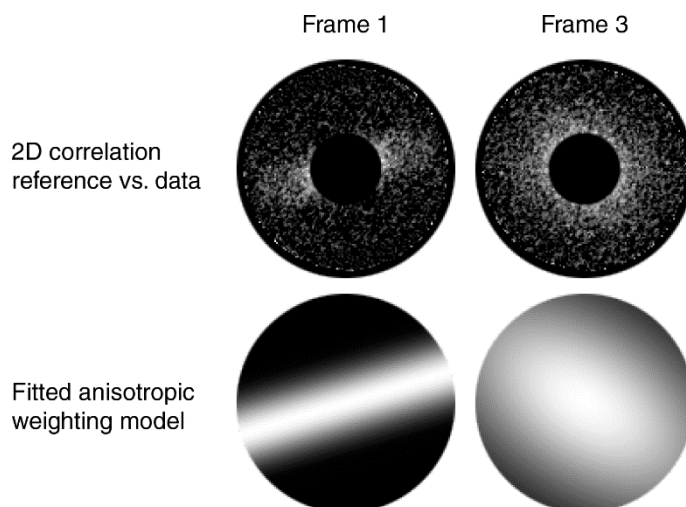


Figure 3.6 | Examples of anisotropic B-factor weighting.

Normalized 2D correlation between reference projections and data, averaged over all particles in a single frame is shown for the 1st and 3rd frame of the same exposure. Values in the low-frequency region are excluded to reduce the value range. The fitted B-factor is highly anisotropic for the 1st frame because of intra-frame motion: 0 Å² and -62 Å² along X and Y, respectively. For the 3rd frame, the fit is much more isotropic due to lack of intra-frame motion, but some high-resolution information is lost to radiation damage: -8 Å² and -10 Å² along X and Y, respectively.

At the end of an optimization iteration, similar to the Fourier Ring Correlation (FRC) approach introduced previously^{39, 87}, *M* calculates the per-Fourier component normalized cross-correlation (NCC) between reference projections and image data. This can be used to empirically optimize exposure- and tilt-dependent data weighting, and reconstruct new half-maps using the updated model, correcting for Ewald sphere curvature⁸⁸. Because the NCC is resolved in 2D, unlike the FRC, anisotropic weights can be fitted to make better use of the first frames, which are often affected by strong, unidirectional motion (**Figure 3.6**). Then, various map metrics, including global, local, and anisotropic resolution, are calculated. Further optimization iterations can be performed to arrive at a denoised or low-pass filtered and sharpened map. Alternatively, 2D particles or sub-tomograms can be extracted and reconstructed from the raw data using the updated alignment information, to be exported to RELION for further, more accurate classification.

3.1.5 Map denoising and local resolution

Instead of using a traditional Fourier Shell Correlation (FSC)-based approach for local resolution estimation⁸⁹, *M* trains a CNN-based denoiser model using a species' half-maps to filter them to local resolution for the next refinement iteration (see Methods). The denoiser applies the noise2noise⁸² training regime to gold-standard⁹⁰ half-map reconstructions obtained at the end of each refinement iteration in *M* by back-projecting extracted images from the original frames or tilts. Because the noise estimation and filtering are done for each half-map independently, no common artifacts are introduced that could be amplified over subsequent refinement iterations. Even without local resolution filtering, the denoising helps to avoid artifacts that may be introduced and amplified when regions of significantly lower resolution are filtered to higher global resolution.

To assess the benefits of *M*'s denoising, we processed the EMPIAR-10288 data set containing the membrane protein cannabinoid receptor 1-G⁹¹ (**Table S3.1**). The 3.0 Å map published with the original study (EMD-0339) showed overfitting artifacts in the lipid bilayer (**Figure 3.7a**). Processing the data with Warp, RELION and *M* led to only slightly improved resolution of 2.9 Å (**Figure 3.7b**) using 149,308 particles (ca. 15% fewer than in the original study). However, the overfitting artifacts were absent in *M*'s final reconstruction (**Figure 3.7a**).

Denoising was also tested on the EMPIAR-10453 tilt series data set containing SARS-CoV-2 virions with spike proteins displayed on the coronavirus surface (**Table S3.1**). The S1 domain of the spike protein is conformationally heterogeneous and has significantly lower resolution than the more stable parts. Processing the data with Warp, RELION and *M* led to a 3.8 Å map (**Figure 3.7c–e**), improving over the originally obtained 4.9 Å⁹². Repeating the refinement in *M* without denoising decreased the global resolution to 4.1 Å and generated visible overfitting artifacts in the S1 domain (**Figure 3.7c, d**). This is in line with improvements recently demonstrated using different approaches to local map filtering that do not rely on conventional FSC-based estimates^{93, 94}.

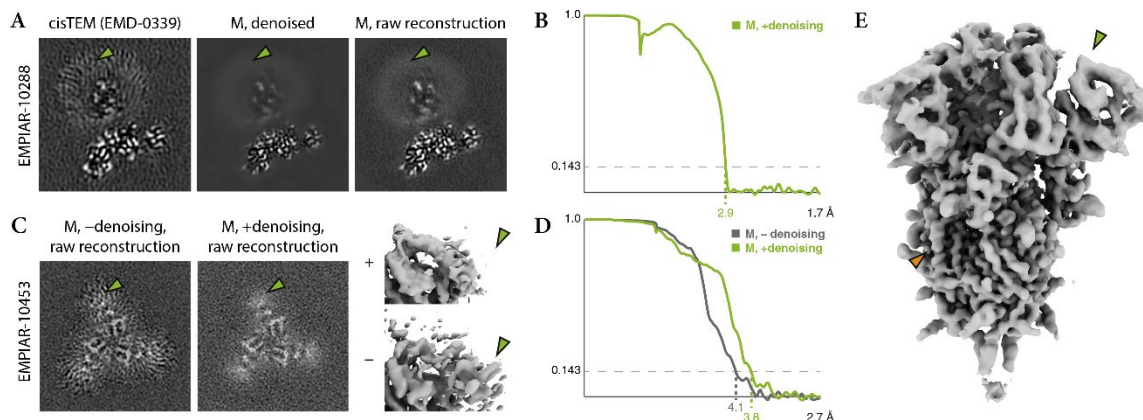


Figure 3.7 | Effects of deep learning-based denoising of reconstructions during refinement.

M trains a denoising model on each species' half-maps after every refinement iteration to filter the maps to local resolution and avoid overfitting artifacts in low-resolution areas, such as lipid nanodiscs or flexible domains.

a) 2D XY slices through 3D reconstructions of the cannabinoid receptor 1-G membrane protein. The original refinement in cistEM (left) introduced artifacts in the highly disordered lipid region (green arrow). The denoised map (middle) and the raw reconstruction before denoising (right) used in the last refinement iteration in *M* are devoid of the artifacts because the denoising filtered and downweighed the low-resolution region.

b) FSC between the half-maps refined in *M*, showing a global resolution of 2.9 Å. A value of 3.0 Å was reported in the original study, with no FSC curve included with the deposited map.

c) 2D XY slices and isosurface renderings of the S1 domain in SARS-CoV-2 spike protein reconstructions. Refinement in *M* without denoising introduced visible artifacts (left, top-right) in the region (green arrows), which had significantly lower resolution than the rest of the protein. Using denoising, the artifacts were avoided (center, bottom-right).

d) FSC between the half-maps refined in *M* with and without denoising, showing an improvement in global resolution from 4.1 Å to 3.8 Å when using denoising.

e) Isosurface rendering of the entire denoised SARS-CoV-2 reconstruction with a global resolution of 3.8 Å. Through the denoising process, the more disordered S1 domain (green arrow) was filtered to lower resolution compared to other parts where side chains are visible (orange arrow).

3.1.6 Contribution of different model parameters to map resolution

We used apoferritin frame and tilt-series data sets, collected from the same grid square under identical conditions (data sets AF-f and Af-t, see Methods), to estimate the contribution of different groups of optimizable parameters to the quality of the reconstructed maps (**Figure 3.8** and **Table S3.1**). For frame series, particles extracted following

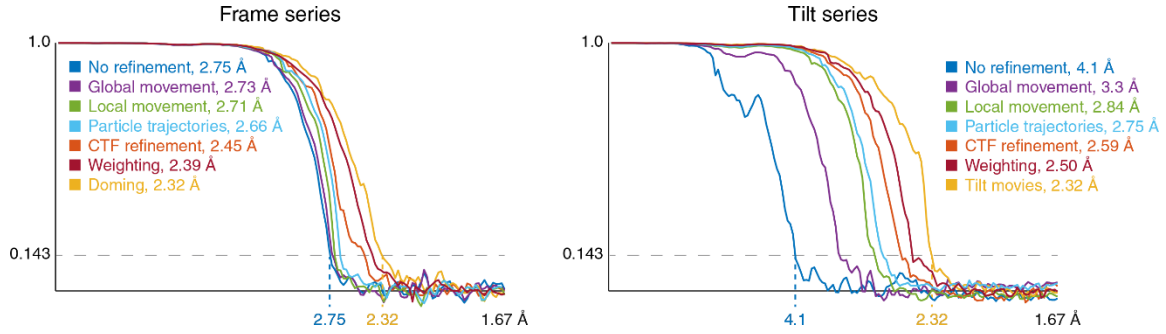


Figure 3.8 | Contributions of individual multi-particle system model components to map resolution.

Fourier shell correlation between half-maps for frame series and tilt series apoferritin data obtained through extending the set of optimizable parameter groups. Starting with the ‘No refinement’ baseline, in top-down order in the legend, a new group of parameters was added, while keeping the previously added groups, and refinement was performed from scratch. The resolution for each step is given in the legend.

reference-free alignment in Warp and refined in RELION (without polishing and CTF refinement) provided a baseline resolution of 2.75 Å, which was then improved by accumulating the following sets of optimizable parameters in *M*: Reference-based global motion alignment improved the resolution to 2.73 Å. Relaxing this constraint to allow local motion alignment improved the resolution to 2.71 Å. Resolving individual particle pose trajectories as a function of exposure led to a resolution of 2.66 Å. Fitting per-particle defocus and per-frame series astigmatism and beam tilt improved the resolution to 2.45 Å. Data-driven anisotropic weight estimation improved the resolution to 2.39 Å. Finally, resolving doming-like motion slightly improved the resolution to 2.32 Å.

For tilt series, sub-tomograms reconstructed following reference-free tilt movie alignment in Warp, patch tracking-based tilt series alignment in IMOD and refinement in RELION provided a baseline resolution of 4.1 Å, which was then improved by accumulating the following optimizations in *M*. First, reference-based global tilt image alignment improved the resolution to 3.3 Å. Relaxing this constraint to allow local image-space warping improved the resolution to 2.84 Å. Resolving individual particle poses as a function of exposure increased the resolution to 2.75 Å. Fitting per-tilt defocus and

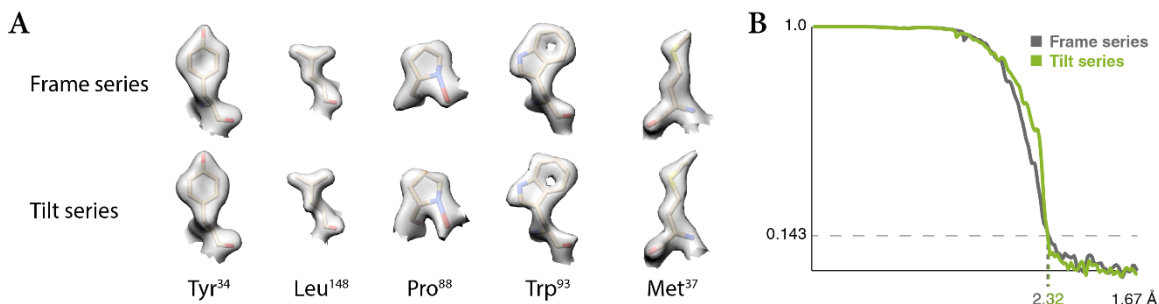


Figure 3.9 | *M* achieves similar resolution for frame series and tilt series data.

Given equal amounts of frame series and tilt series data of similar quality, *M* can achieve identical resolution, closing the gap previously assumed between both data types. This is exemplified on apoferritin data collected in both ways on the same grid.

- a) Representative side chain densities observed in the frame series and tilt series maps.
- b) Comparison between the global FSC curves for each map.

astigmatism, and per-tilt series beam tilt improved the resolution to 2.59 Å. Data-driven anisotropic weight estimation improved the resolution to 2.50 Å. Finally, reference-based tilt movie alignment led to a resolution of 2.32 Å. Volume-space warping was not tested because the particles were arranged in a single 2D layer.

We conclude that accurately registering image-space deformation is essential for obtaining high-resolution maps from frame and tilt series data, whereas modeling other effects leads to smaller improvement that may generally only become significant in the sub-5 Å resolution range. Initial reference-free alignment is significantly less accurate for tilt series than for frame series. However, it is accurate enough to obtain initial reference maps and particle poses that can be further refined in *M*.

3.1.7 Similar resolution obtained from frame and tilt series data

Because tilt series are often associated with lower resolution compared to frame series, we processed both types of data collected from grid holes in close proximity (Af-f and AF-t) to test potential intrinsic limitations of the tilt series data. Given equal amounts of particles, *M* was able to achieve the same resolution with very similar map features (**Figure 3.9**) from either frame series or tilt series data. Thus, collecting data as tilt series does not incur a resolution penalty. However, because tilt series data are still much slower to

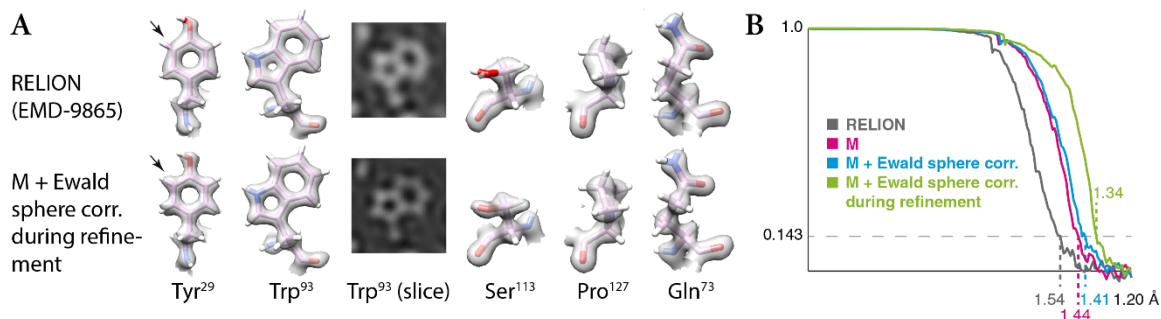


Figure 3.10 | Comparison with RELION on atomic-resolution frame series data.

Atomic-resolution data of apoferritin previously refined with RELION 3.1 to 1.54 Å (EMD-9865) were processed with *M* to achieve a resolution of 1.34 Å, showing that *M*'s image artifact model is suited for very high resolution.

a) Examples of side-chain densities produced by RELION (top) and *M* (bottom), showing cases of improved atomic features such as one of the hydrogens in Tyr²⁹ (black arrow).

b) FSC between the half-maps produced by RELION (grey) and *M* (green), showing a general improvement in resolution through *M*.

acquire⁹⁵ and commonly used for more crowded, thicker samples, we expect maps derived from tilt series data to remain at lower resolution on average.

3.1.8 Comparison with RELION on atomic-resolution frame series data

To assess whether *M* can provide further improvements for frame series data processed with RELION 3.1, we refined a previously published⁹⁶ apoferritin data set (EMPIAR-10248, **Table S3.1**). The data were acquired on a novel JEOL microscope with a cold field emission gun to achieve an atomic resolution of 1.54 Å. Adding *M* to the pipeline improved the resolution to 1.35 Å, revealing densities for hydrogen atoms (**Figure 3.10**). This shows that the image artifact model implemented in *M* can correct data at the highest end of the SPA resolution currently possible. At this high resolution, we were also able to assess the effect of Ewald sphere correction with the single side-band algorithm⁸⁸. Applying it to the reconstruction alone, as would be possible in RELION 3.0, improved the resolution from 1.44 to 1.41 Å. Correcting the particle data and considering the sphere curvature during the multi-particle system refinement, improved the resolution further to 1.34 Å. Coupled with the demonstrated benefits of multi-species refinement and map denoising, this makes *M* a useful addition to the frame series SPA pipeline.

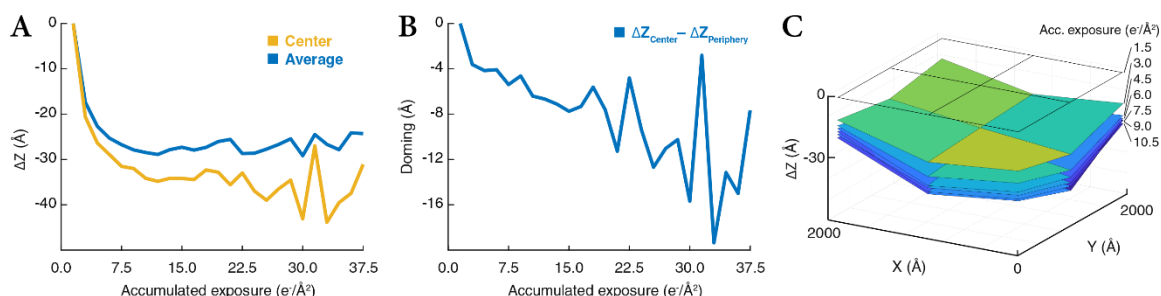


Figure 3.11 | Quantification of the doming effect in frame series data.

Doming models describing per-frame, spatially resolved (3x3 points) defocus offsets fitted during the refinement of atomic-resolution data of apoferritin (EMPIAR-10248) were averaged across the data set, showing significant changes in the CTF during exposure.

a) Defocus change plotted against the accumulated exposure show a fast change in both the central point and the average of the entire field of view's 3x3 points at the beginning of the exposure. After the first 7.5 $e^-/\text{Å}^2$ of exposure, the average change stabilizes, while the central point continues to decrease in defocus.

b) When corrected for global inclination, the difference between the central and peripheral defocus change indicates a steady increase in doming within the field of view as a function of accumulated exposure.

c) Surface rendering of the spatially resolved defocus change for the first 7 frames shows an inclination of the entire field of view as well as a more localized dent in the center.

The high resolution of the data also enabled a detailed analysis of the sample's doming behavior during exposure, as modeled in *M*. The defocus of the entire field of view changed by over -25 Å during the first 7.5 $e^-/\text{Å}^2$ of exposure (**Figure 3.11a**), corresponding to the sample moving away from the electron source. After this initial global decrease, a more localized, steadily increasing bending of the center relatively to the periphery was observed, reaching a difference of ca. -16 Å after 37.5 $e^-/\text{Å}^2$ (**Figure 3.11b, c**). However, because the observed change in the CTF can also be caused by electrostatic lensing effects due to sample charging, further experiments are necessary to investigate the exact nature of doming.

3.1.9 Comparison with other tools for tilt series data refinement

To compare *M*'s reference-based tilt series alignment performance with the previously published EMAN2⁹⁷ and emClarity⁴³ packages, we reprocessed some of the data sets used in the respective publications (**Figure 3.12** and **Table S3.1**). EMAN2 reached a resolution

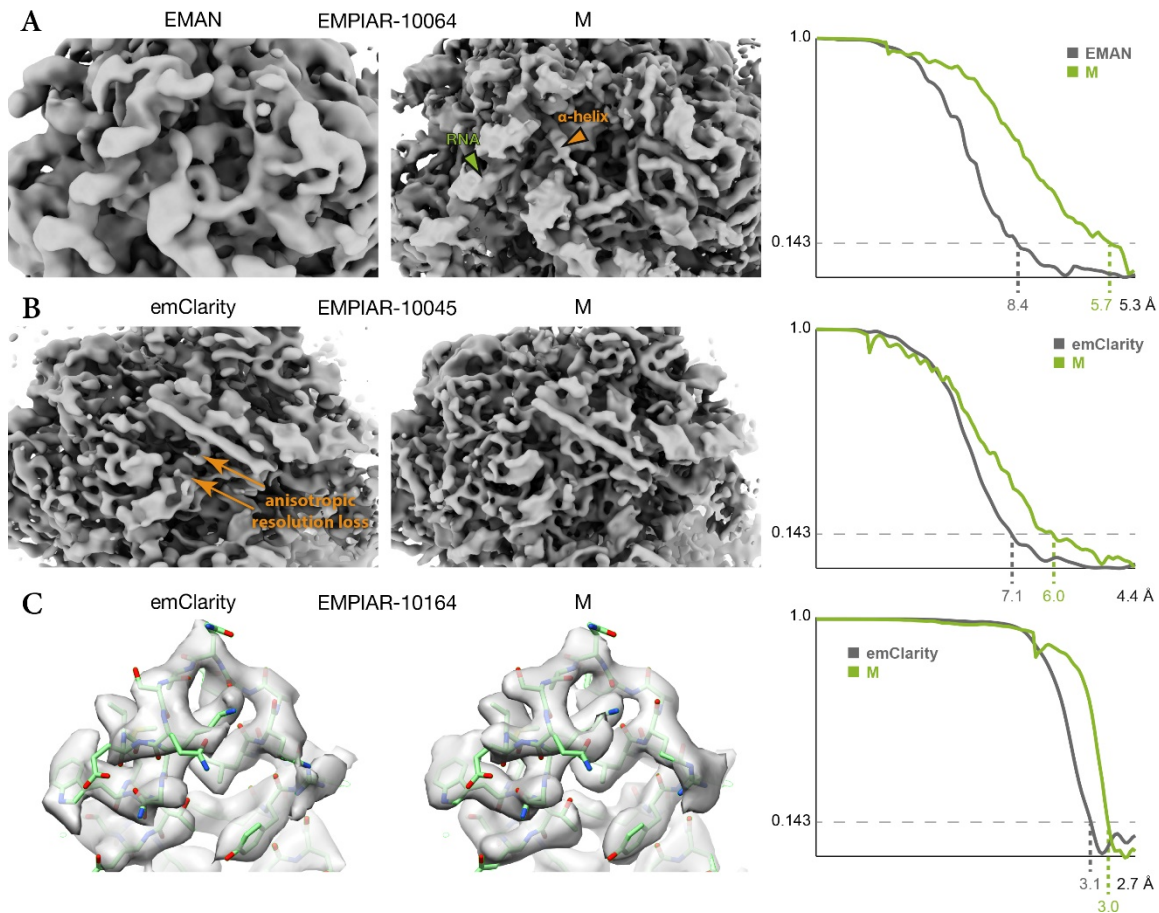


Figure 3.12 | Comparison of maps obtained from published tilt series using *M* or other software.

When applied to data used previously to test the EMAN and emClarity packages, *M* produces maps that compare favorably in terms of resolution and visual features.

a) 80S ribosome data from EMPIAR-10064 were used to benchmark the new tilt series processing in EMAN (EMD-0529). *M* achieved higher resolution, accompanied by visibly better resolved features such as RNA (green arrow) and α -helices (orange arrow).

b) 80S ribosome data from EMPIAR-10045 were used to benchmark emClarity. The originally published map (EMD-8799, not shown) exhibited strong resolution anisotropy. A recently updated map shown here still suffered from resolution anisotropy (“smearing” direction indicated by orange arrows). *M* achieved higher and more isotropic resolution, aiding the map’s interpretability.

c) HIV-1 capsid-SP1 data from EMPIAR-10164 were used to benchmark emClarity. Here, *M* achieved slightly higher resolution using ca. 30% of the particle number used by emClarity. Doubling the number of particles did not increase the resolution.

of 8.4 Å on an *in vitro* 80S ribosome sample (EMPIAR-10064), improving significantly upon a previous 13 Å result⁹⁸. For emClarity, a resolution of 8.6 Å was reported for the same data⁹⁹. *M* improved the resolution to 5.7 Å and resulted in a map that clearly showed secondary structure elements and the helical grooves of the RNA (**Figure 3.12a**). We attribute a significant part of this improvement to *M*'s application of constraints between individual particle tilt images, which is not part of EMAN2.

We further tested *M* on two data sets used in emClarity's benchmarking. The emClarity software reached a resolution of 7.8 Å on purified 80S ribosomes (EMPIAR-10045) in the original publication⁴³, and was later refined to 7.1 Å⁹⁹, improving significantly upon a previous 12.9 Å result³⁸. *M* improved the resolution to 6.0 Å, accompanied by improved resolution isotropy and map features (**Figure 3.12b**). We attribute the improved resolution isotropy to *M*'s denoising-based map filtering approach that learns the optimal filtering empirically, whereas emClarity employs an FSC-based approach that may have to be tuned more conservatively to achieve the desired robustness to artifacts.

It was also reported that emClarity achieved a resolution of 3.1 Å on a thicker sample with a locally high particle density of isolated HIV-1 capsid-SP1 assemblies (EMPIAR-10164), improving upon previous 3.9 Å⁷¹ and 3.4 Å⁶⁵ results. *M* improved the resolution to 3.0 Å, accompanied by local improvements in map quality (**Figure 3.12c**). We attribute the slight improvement of overall resolution in this and the EMPIAR-10045 data sets to *M*'s more accurate deformation model and simultaneous optimization of all parameters, in contrast to emClarity's separate steps for full image alignment (performed in IMOD⁸³) and particle alignment (performed in emClarity). Our results show that *M* can improve over current methods, and achieve higher resolution with various tilt series data sets.

3.1.10 *M* enables the visualization of an antibiotic bound to 70S ribosomes at 3.5 Å in cells

To assess *M*'s performance on *in situ* data in the strictest sense, i.e. in tilt series obtained from intact cells, we used a data set of chloramphenicol-treated *Mycoplasma pneumoniae*¹⁰⁰ (**Figure 3.13a**). *M* was able to resolve the 70S ribosome at 3.5 Å (**Figure**

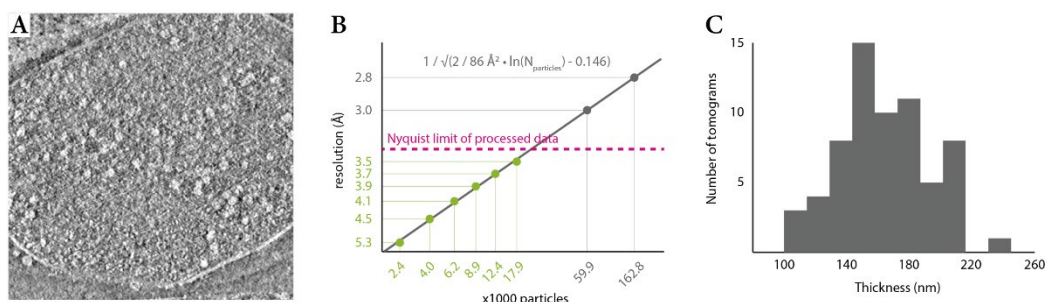


Figure 3.13 | Overview of the *M. pneumoniae* data set.

a) 2D XY slice through an exemplary tomogram.

b) Resolution plotted against the number of particles shows that 5 Å can be obtained with less than 3000 large, asymmetric particles in cells. Extrapolation beyond the Nyquist limit of the data (magenta line) is speculative, but indicates that 3 Å could be surpassed with less than 100,000 particles, given data with higher magnification

c) Histogram of manually measured tomogram thickness values.

3.14a,d) based on 17,890 particles from 65 tomograms, and a B-factor¹⁰¹ of 86 Å² (**Figure 3.13b**). The obtained map exhibited a wide range of local resolution values (**Figure 3.14b, d**). The large 50S ribosomal subunit dominated the alignment and had a higher average resolution than the small 30S subunit, with much of its core reaching the 3.4 Å Nyquist limit of the data. Independent refinement of the 30S and 50S subunits further improved the resolution to 3.7 Å and 3.4 Å, respectively (**Figure 3.13d**). In contrast, processing these data with Warp and RELION alone led to a 10 Å-resolution map of the 70S ribosome (**Figure 3.14c**). *M*'s result constitutes a dramatic improvement compared to the previously used Warp-RELION pipeline, leading to a striking increase in structural detail (**Figure 3.14e**).

The map possesses features typical for this resolution range, such as amino acid side chain stubs and β-sheets with individually resolved β-strands (**Figure 3.14e**). A rigid body fit of an *E. coli* 70S ribosome–chloramphenicol structure (PDB-4v7t) further revealed the presence of a density corresponding to the chloramphenicol molecule at its expected target site (**Figure 3.14f**), marking the first direct visualization of a drug bound to its target inside a cell. The density was absent in a 5.6 Å 70S reconstruction from untreated *M. pneumoniae* cells produced from processing with the older 1.0.6 Warp/*M* versions¹⁰⁰

(**Figure 3.14f**). Therefore, tilt series data acquisition on an intact cellular specimen with a mean thickness of 160 nm (**Figure 3.13c**), in combination with the multi-particle refinement introduced in *M*, can lead to residue-level resolution structures of macromolecules in their native biological context.

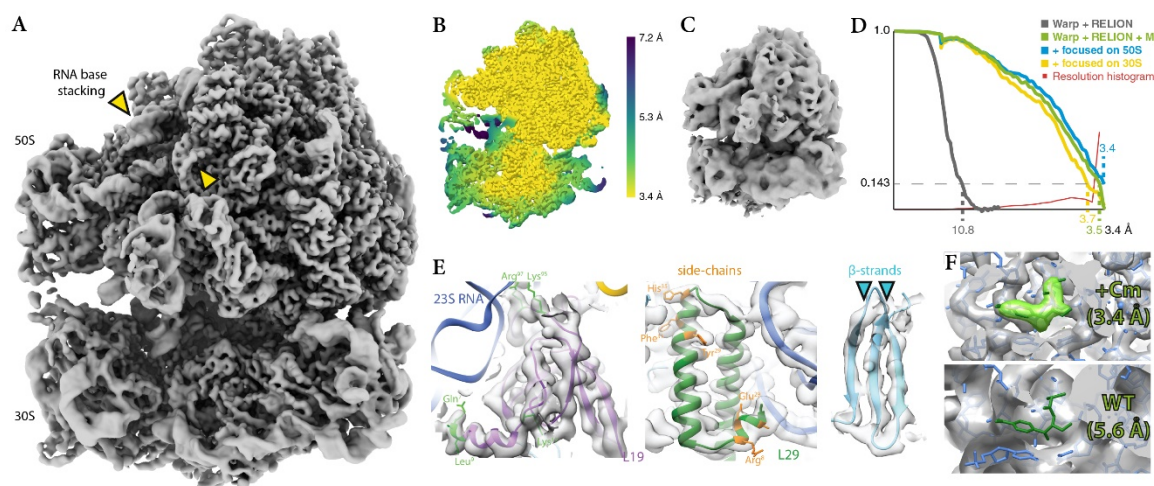


Figure 3.14 | *M. pneumoniae* 70S ribosome-antibiotic map at 3.5 Å refined with the new Warp-RELION-M pipeline.

We applied the Warp-RELION-M pipeline to a tilt series data set of intact cells. The achieved resolution reveals residue-level detail and a bound molecule of the antibiotic chloramphenicol (Cm).

a) Isosurface representation of the 3.5 Å resolution map.

b) Isosurface of the same map colored by local resolution. Despite stalling of the ribosome that is induced by antibiotic binding, residual ratcheting occurs that leads to higher resolution in the large 50S subunit, which dominates the alignment, and lower resolution in the small 30S subunit.

c) Isosurface of a 10.8 Å map derived from the same data set using only Warp and RELION shows the striking increase in detail after refinement with *M*.

d) Comparison between the FSC curves of the 3.5 Å and 10.8 Å map shows the increase in resolution achieved with *M*. Focused refinement of the 30S and 50S subunits further increased their resolution to 3.7 Å and 3.4 Å, respectively. The overlaid local resolution histogram of the 3.5 Å 70S map shows that a significant portion of the map is resolved close to the data's Nyquist limit of 3.4 Å.

e) High-resolution features, such as large amino acid side chains (in green and orange) and well-separated β -strands (cyan arrows), are resolved at a level expected for this resolution range.

f) Atomic model of a Cm-bound 70S ribosome (PDB-4v7t) fitted into the 3.4 Å 50S map (top) shows correspondence of map density (light green) to the Cm molecule (dark green). Fitting the same model into a 5.6 Å 70S ribosome map of untreated *M. pneumoniae* cells (EMD-10683, bottom) does not show any density for Cm, providing a negative control.

3.2 Methods

3.2.1 Data management

M requires data sources initialized based on a Warp project folder. Beside a list of frame/tilt series items, it stores the deformation model to be refined. *M* saves the refined deformation model for each item in the same XML metadata files previously created by Warp. Due to a shared code base, Warp can use the updated model when calculating new frame series averages or tomographic reconstructions. Multiple data sources of either type can be combined in a single population to facilitate the sharing and pooling of valuable in situ data that can contribute to far more than one project, but do not contain enough data for any single project on their own. To account for minor pixel size miscalibrations between different microscopes, the pixel size can be refined alongside other parameters in *M*.

A species is initialized from the refinement results of RELION or other compatible software, taking the unfiltered half-maps, a mask, and the particle coordinates and poses (i.e. translations and rotations) as a starting point. The state of a species after each refinement iteration comprises the reconstructed half-maps, the weights of the trained denoising model, various filtered and sharpened maps, a denoised map, and a list of particle coordinates and poses with multiple temporal sampling points if desired. The particles reference their data source items by their data hash to avoid naming conflicts between different data sources.

To enable multiple users to collaborate and pool their results, *M* tracks precisely the chain of refinements and other operations on data. After each refinement iteration, a “commit” is generated to save the new state. Similar to version-control systems like Git¹⁰², the commit’s hash is based on the exact state of the system committed. The hash of each data source item is calculated from the raw data, the refined deformation and imaging models, and the hashes of all species used for their refinement. The hash of each species is calculated based on the half-maps, the weights of the denoising model, the particle coordinates and poses, and the hashes of all data source items contributing information. The

hashes can be used to verify a graph representing all steps that led to a particular state of a data source or species. Similar to the “pull request” mechanism in Git, species can be added to a population taking into account potential physical collisions with existing particles. This enables the maintenance of a centralized population repository from which multiple users can obtain pre-aligned data sources, identify new particle species or re-classify existing particles into more states, and contribute the results back to the repository.

3.2.2 Deformation model

For frame series data, deformation of the multi-particle system is modeled in the XY plane only, with a pyramid (**Figure 3.3**) of cubic spline grids⁸⁴ $G_{F,j}(\delta, i)$ (where j is the index within the pyramid, δ is the spatial interpolation coordinate, and i is the temporal interpolation coordinate) going from high temporal/low spatial to low temporal/high spatial resolution. This accounts for the fast-changing, global stage movement, and the slowly-developing, local BIM. Furthermore, translation and rotation of individual particles as a function of exposure can be modeled with 2–3 control points depending on the particle size and overall exposure.

The model for tilt series data is more complex, owing to the higher potential for perturbations in the system between individual tilt exposures. As the mechanical rotation of the microscope stage and the estimated orientation of the tilt axis are imperfect, the assumed stage orientation can be randomly off in every tilt. M thus refines an independent set of stage rotation angle corrections ω_i for every tilt i . These corrections only affect the particle orientations to avoid redundancy, as the induced changes in the projected particle positions can be fully modeled by a deformation grid that must already be employed for other purposes.

Similarly, stage translation varies randomly between individual tilts. BIM patterns can be very different across adjacent tilt images as additional exposures are taken for focusing and tracking in-between. Particle positions can further deviate due to other imaging artifacts, such as wrongly calibrated magnification anisotropy²². M employs an “image warp”

grid of cubic splines G_{TI} with a spatial resolution of 3–5 in X and Y and per-tilt temporal resolution to model these geometric displacements in image space collectively. Furthermore, *in vitro* and *in situ* sample types for which tilt series are commonly used contain multiple overlapping layers of particles. Some deformations of densely filled volumes, such as shearing, or bending in the Z dimension when viewed at a high tilt angle, cannot be modeled accurately by XY translations in image space. M employs an additional “volume warp” grid G_{TV} , implemented as a 4D grid of control points with quadrilinear interpolation between them that is anchored in volume space rather than image space. Hence it rotates with the sample and can model slow, continuous deformation that affects the particles’ projected positions in image space. As with frame series data, per-particle translation and rotation as a function of exposure is also modeled for tilt series.

Finally, a single tilt image exposure is usually fractionated in multiple frames, making it a tilt movie. At 1–3 $e^-/\text{\AA}^2$, the exposure in a single tilt movie is usually short, but still requires additional modeling to compensate motion. M parametrizes the XY translation as a combination of a grid with no spatial and per-frame temporal resolution, and a grid with a spatial resolution of 3x3 and a temporal resolution of 3. Stage and particle orientations are assumed to remain constant throughout a tilt movie, as the biggest beam-induced changes have been shown to occur in the very beginning of each of the short exposures²⁶. Overall, the number of parameters for tilt series is larger than for frame series, requiring a higher particle density to achieve equivalent accuracy.

3.2.3 Imaging model

The ability to model imaging conditions such as defocus, astigmatism, magnification or higher-order aberrations is equally important for obtaining high-resolution reconstructions. Frame and tilt series offer different advantages for refining some of these parameters.

For particles in frame series data, the Z coordinate and thus the relative offset from the global defocus of the micrograph is unknown. Although local defocus estimation based on amplitude spectrum fitting has been shown to increase resolution⁸⁴, reference-based

refinement of per-particle defocus can lead to a further increase in resolution³³. M refines per-particle defocus and a per-series astigmatism for frame series, assuming constant values throughout the series.

Tilt series, on the other hand, provide accurate Z coordinates for all particles. However, the initial amplitude spectra-based global defocus estimates for each tilt have lower accuracy due to very short exposures, and cannot be assumed to remain constant throughout the series due to stage movement and refocusing. Furthermore, these estimates can be biased by contrast-rich objects that are not the particles of interest, such as a carbon film below or above the particles, or the platinum coating layer for FIB-thinned samples¹⁰³. The astigmatism can also change between tilts due to fluctuating electron optics. M refines per-tilt defocus and astigmatism for tilt series, and calculates per-particle tilt CTFs based on these values and the Z coordinate of a particle's position transformed according to the fitted stage orientation. Particles in tilt series can potentially have more accurate defocus values because the number of parameters that can be fitted scales with the number of tilts or particles for tilt or frame series, respectively. In many cases the number of tilts will be significantly lower than the number of particles.

In both frame and tilt series, M also models per-series anisotropic magnification and higher-order optical aberrations. Refinement of a global set of Zernike polynomials representing the aberrations based on a 2D phase residual image calculated from all particles in a data set has been shown to improve the resolution significantly for slightly misaligned microscopes¹⁴. Within individual tilt series, too, beam tilt can vary as it is applied to compensate stage misalignments during tracking. Unfortunately, the signal in individual tilts is insufficient for accurate beam tilt estimation, and such an option is not implemented in M .

3.2.4 Optimization procedure

M seeks to maximize the following target function M , which is essentially a weighted, normalized cross-correlation between all particle images and the corresponding reference projections:

$$M = \frac{\sum_s \sum_p \sum_i A_{s,p,i} \cdot B_{s,p,i}}{\sqrt{\sum_s \sum_p \sum_i |A_{s,p,i}|^2 \cdot \sum_s \sum_p \sum_i |B_{s,p,i}|^2}},$$

$$A_{s,p,i} = W_i \cdot P(s, \Theta_{p,i}, \tau),$$

$$B_{s,p,i} = T \cdot \text{FT}(\text{FT}^{-1}(W_i \cdot \text{CTF}(i, \Lambda_{p,i}) \cdot AS_i^{-1} \cdot I(i, \Lambda_{p,i})) \cdot D(d_s)),$$

where s is a particle species, p is a particle of that species, and i is the index of a frame or tilt in a series; \cdot denotes the dot product between two complex vectors, where the complex numbers are treated as pairs of scalars; $|\dots|$ denotes the L_2 norm; W is the anisotropic exposure- and tilt angle-dependent amplitude weighting of frame or tilt i ; P is a projection operator in Fourier space sampling a central slice of the volume of species s at orientation Θ , taking into account the anisotropic scaling τ , bent to account for the Ewald sphere curvature determined by the species' diameter; \cdot denotes scalar multiplication; T is the complex-valued beam tilt compensation; FT denotes the discrete Fourier transform; CTF is the real-valued CTF taking into account the defocus at position Λ and the astigmatism in frame or tilt i ; AS is the real-valued, rotational average over the amplitude spectra of all particle images of all species extracted from tilt i or the average of all aligned frames, used for spectrum whitening, scaled and cropped to the respective species size and resolution; I is the FT of a particle image extracted from frame or tilt i at position δ , cropped to the respective species resolution; D is a soft circular mask with particle diameter d .

Similar target functions in previous literature used $P \cdot \text{CTF}$ to model the contents of I ^{4, 86}. However, in M 's implementation I is pre-multiplied by CTF to avoid CTF aliasing despite using small particle windows. This change does not affect the numerator part of M due to the associativity of complex number multiplication; its impact on the denominator part of M does not affect the achieved resolution in any way. It also avoids the additional memory footprint of storing pre-calculated CTFs, or the computational overhead of calculating them on-the-fly.

M can consider the Ewald sphere curvature during refinement if this is made necessary by a large species and/or high resolution¹⁰⁴. In this case 2 copies of $CTF \cdot I$ are prepared using the single side-band algorithm⁸⁸: $CTF_P \cdot I$ and $CTF_Q \cdot I$. To calculate the cost function, one is correlated with a bent central slice P , and the other with a central slice bent in the opposite direction. The resulting cost functions M_P and M_Q are then added. As with previous implementations³³, the absolute handedness for the correction must be provided by the user.

For frame series, the position and orientation of particle p in frame i are calculated as:

$$\Lambda_{p,i} = \lambda_p(i) + \sum_j G_{OF,j}(\lambda_p(i), i) + \sum_j G_{F,j}(\lambda_p(i), i) + Z_p ,$$

$$\Theta_{p,i} = \theta_p(i) ,$$

where λ is the value of the refined particle position trajectory interpolated at the accumulated exposure of frame i ; G_{OF} is a deformation grid pyramid produced by Warp's original reference-free alignment that is not altered in M refinement; G_F is a deformation grid pyramid that is refined in M ; Z is the refined defocus value of particle p that is added as the Z coordinate to its position; θ is the value of the refined particle orientation trajectory interpolated at the accumulated exposure of frame i .

For tilt series, the position and orientation of particle p in tilt i are calculated as:

$$\Lambda_{p,i} = R(\Omega_i) \cdot (\lambda_p(i) + G_{TV}(\lambda_p(i), i) - C_V) + C_i + G_{TI}(\lambda_p(i), i) + Z_i ,$$

$$\Theta_{p,i} = R^{-1} \left(R_{XYZ}(\omega_i) \cdot R(\Omega_i) \cdot R(\theta_p(i)) \right) ,$$

where R and R_{XYZ} construct a rotation matrix based on a set of Euler and XYZ angles, respectively, and R^{-1} calculates a set of Euler angles based on a rotation matrix; C_V is the center of the volume in which the multi-particle system is anchored, and C_i is the center of the full tilt image; Z_i is the refined defocus value of tilt i that is added to the Z coordinate of the transformed particle position; Ω is the stage orientation determined in the

initial, reference-free tilt series alignment that is not altered in M refinement; \cdot denotes matrix multiplication here.

For frames in tilt movie i , the position of particle p in frame k is calculated as:

$$\Lambda_{p,k} = \Lambda_{p,i} + \sum_j G_{OF,i,j}(\Lambda_{p,i}, k) + \sum_j G_{TF,i,j}(\Lambda_{p,i}, k) ,$$

where G_{OF} is the deformation grid pyramid produced by Warp's original reference-free alignment of the tilt movie that is not altered in M refinement; G_{TF} is a deformation grid pyramid for the tilt movie that is refined in M .

Due to the very large number of parameters, M employs L-BFGS⁶⁶ to perform almost all of the optimization. Only the initial defocus search is done exhaustively over a limited range to avoid getting trapped in a local optimum because of the quickly oscillating nature of the CTF. Every L-BFGS search iteration requires the calculation of a partial derivative of the target function with respect to each optimizable parameter. Reevaluating M twice per parameter to compute the gradient with the central differences numerical scheme would be very computationally expensive. Like Warp, M takes a computational shortcut for most of the parameters.

Before optimization starts, M calculates the partial derivatives of the X and Y components of all $\Lambda_{p,i}$ with respect to all warping grid parameters and all control points of a particle's position trajectory that affect them. Similarly, the partial derivatives of the individual Euler angle components of all $\Theta_{p,i}$ with respect to all stage angle correction parameters and all control points of a particle's orientation trajectory are calculated. As each parameter influences only a small fraction of particle frames or tilts, most of the derivatives are 0. They are excluded from the precalculated lists to avoid unnecessary computation. Then, during optimization, once per search iteration, the partial derivative of $(A * B) / \sqrt{|A|^2 |B|^2}$ for each particle frame or tilt is calculated with respect to X, Y and the Euler angles. This amounts to evaluating M 10 times. A useful approximation for the derivative for each parameter η can then be calculated as follows:

$$\frac{\partial M}{\partial \eta} = \frac{\sum_s \sum_p \sum_i \sum_\alpha \frac{\partial \left(A_{s,p,i} \cdot B_{s,p,i} / \sqrt{|A_{s,p,i}|^2 |B_{s,p,i}|^2} \right)}{\partial \alpha} \cdot K_{s,p,i} \cdot |A_{s,p,i}| \cdot |B_{s,p,i}|}{\sum_s \sum_p \sum_i \sum_\alpha |A_{s,p,i}| \cdot |B_{s,p,i}|},$$

$$K_{s,p,i} = \frac{\partial (\Lambda_{p,i} \parallel \Theta_{p,i})_\alpha}{\partial \eta},$$

where $\alpha \in \{x, y, \phi, \vartheta, \psi\}$, i. e. one of the translation axes or Euler angles; \parallel denotes the concatenation of two tuples; $(\dots)_\alpha$ denotes the selection of component α from a tuple.

The deformation parameters make up the bulk of all parameters. Parameters such as absolute magnification and beam tilt do not benefit from the same shortcut and their derivatives must be calculated independently with the central differences scheme. The CTF-related parameters are few, but the calculation of their derivatives is especially expensive because it requires the particles to be reextracted at an aliasing-free size, pre-multiplied by the altered CTF, and cropped to refinement size – all involving expensive FT steps. *M* calculates the values of *M* by adding up the results from small batches of particles. This allows the cost of the first FT at aliasing-free size to be amortized over all optimizable CTF parameters, as its result is reused for all subsequent calculations. The gradients for all per-particle or per-tilt defocus and astigmatism parameters can all be calculated in the same pass as each of them affects only one particle or tilt.

If defocus is to be optimized, an iterative grid search can be executed before the L-BFGS optimization starts. The search runs for 5 iterations. For the first iteration, a range of ± 300 nm around the current values is sampled in 10 nm steps. For each subsequent iteration, the search step is halved, and a range of \pm the new search step around the 2 best values for each particle or tilt from the previous iteration is sampled.

3.2.5 Memory footprint considerations

Traditional SPA refinement treats every particle as an isolated entity, thus requiring no more than one particle to be held in memory at any given time if parallelization is not considered. A multi-particle approach, however, needs to rapidly evaluate the state of the entire multi-particle system during refinement. The particle frame/tilt series need to

be stored in memory because re-extracting and reprocessing them for every evaluation would be too inefficient. While an *in vitro* sample usually contains a single layer of proteins with up to 1000–2000 particles in a field of view, a densely packed *in situ* volume has the potential to contribute tens of thousands of particles to refinement if enough species can be identified. The image size is selected to be twice the particle diameter to account for signal delocalization and interpolation artifacts, leading to significant overlap even in the single-layer case. At high refinement resolution, the memory requirements of all extracted particle frame/tilt series in a system can vastly exceed those of the original data, rising to tens or even hundreds of gigabytes.

Although *M* uses GPUs for acceleration wherever possible, currently available consumer-level cards offer up to 12 GB, which would be insufficient in many cases. Therefore, the extracted particle frame/tilt series are held in “pinned” (i.e. page-locked) CPU memory where they can be transparently accessed by the GPU. Despite the low bandwidth of CPU–GPU memory transfers, the GPU does not experience a significant performance penalty when correlating them to reference projections. This is because the particle data accesses are sequential and highly coalesced, whereas the creation of reference projections on-the-fly accesses the GPU memory randomly, creating significant overhead. As faster CPU–GPU interfaces are being developed, the penalty should become more negligible in the future.

Still, memory requirements can become too high even for CPU memory. To reduce the footprint, *M* exploits the varying information content of frames/tilts over the course of a series. As sample damage from radiation is accounted for by applying a Gaussian (“B-factor”) weighting function in Fourier space^{7, 33}, the contribution of higher-frequency components becomes negligible at high exposure. *M* crops extracted particle images in Fourier space to a resolution that corresponds to the weighting function value falling below 0.25, resulting in considerable space savings once high resolution is reached. Assuming an increase in the weighting B-factor of 4 \AA^2 per $1 \text{ e}^- / \text{ \AA}^2$ of accumulated exposure, the maximum useful frequency at exposure d is $f_{max} = \sqrt{\ln(4)/d}$, and the image size m

scales with a factor of $\min(1, f_{\max}/f_{\text{refine}})$. Thus, the upper bound for memory consumption in case of low refinement resolution and/or low overall exposure is $O(m^2 d)$, while the lower bound is $\Omega(m^2 \ln(d))$ in case of high refinement resolution and/or high overall exposure.

3.2.6 Avoiding CTF aliasing

Cryo-EM data of thin biological specimens are usually acquired at defocus to achieve phase contrast. In the absence of a phase plate device, and often in the case of *in situ* tomography, defocus values can exceed 4 μm to enable better visual interpretation of the raw data. Higher defocus results in stronger delocalization of the signal in real space, as reflected by faster oscillations of the CTF in Fourier space. As the CTF oscillates between -1 and 1, combining signals with different defoci would result in an average value of 0 at higher spatial frequencies. Thus, a phase shift of π must be applied to frequency components modulated by negative CTF values prior to averaging. Furthermore, it is desirable to compute the reconstruction as a weighted average, using the CTF for the weighting. Multiplying the FT of a particle image by the corresponding real-valued CTF achieves both goals.

Current SPA packages advise the user to select the particle box size as 1.5–2 the particle diameter to account for Fourier-space interpolation artifacts, not considering the image defocus. When an image is cropped around a particle, the Fourier-space modulation pattern becomes band-limited to the new window size. If CTF oscillations are too fast to be resolved, the band-limited values for the amplitudes of the corresponding frequency components will converge to 0. Even worse, the analytical 2D CTF model used in refinement and reconstruction is not band-limited, and contains solely aliasing artifacts past the Fourier-space Nyquist frequency instead of converging to 0. This can put a hard limit on the achievable resolution for small particles and those acquired at high defocus that is independent of the actual data quality.

This problem can be mitigated by selecting a box size large enough to avoid CTF aliasing⁶⁸ at the highest defocus value in a data set. However, the required size m can exceed 1000

px at high resolution or defocus, significantly slowing down refinement algorithms whose complexity and memory footprint are $O(m^2)$ and $O(m^3)$, respectively. This increase can be entirely avoided by pre-multiplying particle images by the CTF at an aliasing-free size, and cropping them to a smaller size for refinement or reconstruction. As the modulation pattern is CTF^2 after pre-multiplication, the band-limited oscillations will converge to 0.5 instead of 0. The 2D CTF model used in refinement and reconstruction must be similarly band-limited to match the data. As M operates on all particles of an entire frame/tilt series at a time and extracts the particle images on-the-fly, such considerations are made automatically for the currently needed resolution.

The minimum box size needed for CTF correction at a given resolution is dictated by the maximum oscillation rate of the CTF within the available spatial frequency range. This is not necessarily the oscillation rate at the highest spatial frequency as φ is not a monotonic function: A combination of low underfocus and high C_s will cause the oscillations to slow down significantly and accelerate again at higher spatial frequencies. The oscillation rate can be calculated as the first derivative of φ . In practice, it is easier to evaluate $d\varphi/dk$ numerically within the relevant range of spatial frequencies to find its maximum absolute value. To fully resolve the oscillation, one period must be rasterized onto at least 2 pixels, i.e. the window size must be chosen such that $\max(d\varphi/dk) = 2\pi/2px$. While this guarantees a fully resolved CTF in 1D, a CTF rasterized on a Cartesian 2D grid has an anisotropic sampling rate. At its lowest, i.e. along the diagonals, it requires $\sqrt{2}$ the sampling rate of the 1D case.

Before particle extraction, the size padding factor at which the images will be pre-multiplied by the CTF has to be determined, taking into consideration the maximum defocus value expected in a frame/tilt series, and the expected maximum resolution. During refinement, the latter is set to the refinement resolution. For the final reconstruction, it is set to 1.25x the current global resolution. Particles are extracted using the calculated minimum box size (or twice the particle diameter in case that value is larger), and pre-multiplied by the CTF in Fourier space. Then the inverse FT (IFT) is applied, the particles

are cropped to the refinement or reconstruction size in real space, and transformed back to Fourier space for refinement. The band-limited CTF² model is prepared by simulating the function at the same aliasing-free size in Fourier space, cropping its IFT in real space, and taking the real components of the result's FT.

3.2.7 Data-driven weighting

To account for radiation damage as a function of accumulated exposure, or increasing sample thickness as a function of the stage orientation, several heuristics and empirical approaches have been proposed^{7, 33, 38}. By default, *M* adopts the heuristic introduced in RELION 1.4³⁸. The B-factor is increased by 4 Å² per 1 e⁻/Å² of exposure, and each tilt is weighted as $\cos \vartheta$. Once high resolution is reached, the weights can be estimated empirically using a reference correlation-based approach similar to the one introduced in RELION 3.0³³.

In a departure from RELION's scheme, the normalized correlation (NC) is calculated between particle images and reference projections at the end of a refinement iteration are not combined across the entire data set. It is kept as a 2D image to enable the fitting of anisotropic weights rather than averaging rotationally. The correlation data can then be recombined in different ways to calculate different kinds of weights. Furthermore, because *M* supports the refinement of multiple species with different resolution, the per-species correlation vectors for each frame or tilt need to be combined. This is done by weighting each one by the FSC calculated between the half-maps of the respective species. This produces a set of vectors $NC_{d,i,k}$, where *d* is the series, *i* is the frame or tilt, and, optionally, *k* is the tilt movie frame.

The procedure then iteratively calculates \overline{NC} as:

$$\overline{NC} = \frac{\sum_d \sum_i \sum_k NC_{d,i,k} \cdot G(B_d + B_i + B_k) \cdot W_d \cdot W_i \cdot W_k \cdot \overline{CTF}_{d,i}}{\sum_d \sum_i \sum_k G(B_d + B_i + B_k) \cdot W_d \cdot W_i \cdot W_k \cdot \overline{CTF}_{d,i}},$$

and optimizes the weighting parameters to minimize the following cost function:

$$C = \sum_d \sum_i \sum_k |NC_{d,i,k} - \overline{NC} \cdot G(B_d + B_i + B_k) \cdot W_d \cdot W_i \cdot W_k|,$$

where \cdot denotes scalar multiplication; G is an anisotropic 2D Gaussian B-factor weighting function; B is a vector describing the B-factor along the X and Y axes, and their rotation; W is a scalar weight; \overline{CTF} is the weighted average of all particle CTFs in one frame or tilt. The B-factors in each group are constrained such that the highest value in a group is set to 0.

In this default formulation, the weighting scheme allows to assign separate weights not only to individual frames/tilts, but also to weight the contribution of an entire series. For data with high particle density this scheme can be extended to assign different weights to frames/tilts of each individual series. Anisotropic B-factors improve the weighting of frames with significant intra-frame motion (**Figure 3.6**). Combined with per-series, per-frame weighting, such granularity allows to rescue more information from the first few frames of an exposure if parts of them are less affected by BIM.

3.2.8 Map reconstruction

Previous refinement packages took two different approaches to map reconstruction from frame and tilt series data. For frame series, weighted averages were prepared either directly from the initial, reference-free alignments, or based on a “polishing” procedure³³. These 2D averages were then weighted based on a 2D CTF model and a spectral signal-to-noise ratio (SSNR) term⁴, and back-projected to obtain the reconstruction. For tilt series, the algorithms operated on intermediate per-particle 3D reconstructions (‘sub-tomograms’) with fixed translational and rotational offsets between individual tilt images. These 3D sub-tomograms were then weighted based on a 3D CTF model³⁸ and an SSNR term, and back-projected to obtain the reconstruction.

M seeks to unify the handling of both types of data and uses the original, non-interpolated 2D data at every step, including reconstruction. For tilt series, this approach avoids any artifacts from intermediate interpolation and reconstruction steps. For frame series, the requirement for identical orientation of all particle frames no longer exists as they are not averaged in 2D, enabling the modeling of particle orientation as a function of exposure. Only for individual tilt movie frames a shortcut is taken to save memory and

computation, and they are pre-averaged in 2D using the approach described for Warp⁸⁴ after a separate multi-particle refinement of the respective tilt movie.

Thus, for the reconstruction, individual particle frames or tilts are weighted by an exposure-dependent function to account for radiation damage, and an aliasing-free 2D CTF model (see previous section) that incorporates the exact defocus and astigmatism values for that position and frame/tilt. The weighted data are then back-projected through Fourier space summation, accounting for Ewald sphere curvature. The reconstruction is finalized by dividing the summed data component by the summed weights component⁴.

3.2.9 Map denoising

Reconstructions of biological specimens derived from cryo-EM data rarely have homogeneous resolution throughout all parts of the macromolecule. Using a map filtered to its global resolution for particle alignment can have detrimental effects. Poorly resolved regions, such as floppy protein domains or the lipid bilayer around transmembrane domains, will make the alignment worse by adding noise to reference projections below the refinement resolution. In the case of fully independent half-maps⁹⁰, the noise patterns that the particles will be aligned against are independent, and amplifying them over several iterations only has the potential of making the resolution worse. In the case of refinement with merged half-maps⁸⁶, where overfitting is avoided by limiting the refinement resolution, the poorly resolved regions may be well below that limit, leading to a common, overfitted noise pattern in both half-maps.

Past attempts at filtering maps based on local resolution estimates for refinement^{28, 105} applied FSC-based approaches⁸⁹ to estimate the local resolution and performed the filtering in the Fourier domain. As only one set of estimates can be made based on one pair of half-maps, any spurious patterns in the estimated values will be introduced into both half-maps when the filtering is performed. The locality and accuracy of the estimates depends on the window size⁸⁹. A smaller window increases locality at the expense of accuracy. Once introduced, the noise pattern can become amplified over multiple iterations, leading to overestimated local resolution and phantom features that can be

misinterpreted. More advanced regularization schemes have been proposed^{93, 94} since to deal with this problem.

M implements a new approach to map filtering that uses neural network-based denoising. The recently proposed noise2noise training principle⁸² allows the training of differentiable denoiser models without a noise-free ground truth, using only two independently noisy observations. It has been successfully applied to micrograph⁸⁴ and tomogram^{52, 84} denoising. The implementation in *M* utilizes gold-standard⁹⁰ half-map reconstructions, which represent another obvious case of two independently noisy observations of the same signal, and are interchangeably used as input and target in training. The reconstructions are obtained at the end of each refinement iteration in *M* by back-projecting extracted images from the original frames or tilts, using the particle half-sets carried over from RELION at the beginning of the workflow. We find that a denoiser trained on one pair of half-maps not only matches closely the result of conventional global resolution filtering when applied to maps with homogeneous resolution, but also provides locally smooth, artifact-free local resolution filtering. As such models can train on and denoise sets of micrographs or tomograms with different defocus values and thus different noise models, they can also recognize and adapt to different noise levels within the same reconstruction. In another important departure from FSC-based methods, the denoising step is applied to the half-maps independently and the denoiser sees only one of them at a time. Thus, even if some spurious pattern is introduced as part of the denoising, it is independent between the half-maps.

The neural network architecture is identical to the one used for tomogram denoising in Warp. A separate denoising model is maintained for every species, and trained only on the respective pair of half-maps. The model is initialized with random values and trained for 800 iterations upon the creation of a new species. It is later retrained for another 800 iterations after every refinement. Spectrum whitening is applied to the maps before training to restore high-frequency amplitudes⁸⁶, similar to B-factor-based sharpening¹⁰¹. During training, 64^3 px volumes are extracted from both maps at the same random

position and orientation, and presented to the network as input and output in mini-batches of 3. The random orientations make sure the network learns the noise model rather than merely learning the average map. The learning rate for the Adam optimizer is exponentially decreased from 10^{-3} to 10^{-5} throughout the training. For the denoising of each half-map, the map is partitioned in 64^3 px windows overlapping by 24 px, denoised, and the results from each window are inserted into the output volume. Regardless of regions with above-average resolution being potentially present, the refinement resolution is set conservatively to the global map resolution. In addition to the two half-maps for refinement, a denoised average map is also prepared by applying the same denoising model to the average of the spectrum-whitened half-maps.

3.2.10 Assessment of map denoising

Frame series data were downloaded for the EMPIAR-10288 entry (**Figure 3.7a, b**). Frame alignment and local CTF estimation were performed in Warp with a spatial resolution of 5x5. 1,033,994 particles were picked with a retrained BoxNet model in Warp and exported at 1.5 Å/px. 2D classification, 3D classification and refinement were performed in RELION using EMD-0339 as the initial reference. 149,328 particles corresponding to the best 3D class were imported in *M*. The particle poses were given a temporal resolution of 2, the deformation grid resolution was set to 2x2, and refinement of all parameters was performed for 5 iterations (**Table S3.1**). Data-driven weight estimation was performed to assign unique weights to every frame index.

Pre-aligned tilt movies were downloaded for the EMPIAR-10453 entry (**Figure 3.7c, d**). Gold fiducials were picked with BoxNet in Warp, and fiducial-based tilt series alignment was performed in IMOD. Tilt series CTF estimation and reconstruction of full tomograms at 12 Å/px was performed in Warp. A binary classifier based on a 3D CNN (in development, not part of Warp and *M*) was trained using 5 manually segmented tomograms to segment the SARS-CoV-2 virions. Another 3D CNN-based binary classifier was trained on manually picked spike protein positions in 7 tomograms. Automatically picked spike protein positions were cross-referenced with the segmented virions to remove particles

further away than 200 Å, obtaining 38,742 particles. Sub-tomograms were reconstructed at 5 Å/px for refinement in RELION. After *ab initio* map generation, 3D refinement was performed, reaching the 10 Å Nyquist limit. The results were imported in M, where a 1x1x41 image warping grid and particle poses were optimized for 2 iterations. Sub-tomograms were reconstructed at 5 Å/px using the improved alignments, and subjected to classification into 4 classes in RELION. 22,998 particles from 2 classes showing the spike trimer were imported in M, where a 3x2x41 image warping grid, CTF, and particle poses were optimized for 4 iterations with C3 symmetry (**Table S3.1**). For the comparison, the refinement procedure was modified to omit the denoising step. Refinement was then restarted at 10 Å and performed for 5 iterations using the same settings.

3.2.11 Acquisition of apoferritin benchmark data

To compare the resolution achievable with frame and tilt series data and assess individual algorithms implemented in M, we acquired two data sets of human heavy-chain apoferritin: AF-f (frame series) and AF-t (tilt series). To make sure that any observed differences came from data type and processing strategies rather than local variance in sample quality, neighboring holes within the same grid square were used for both data sets.

The apoferritin plasmid and purification protocol were kindly provided by Louise Fairall and Christos Savva from the Midlands Regional Cryo-EM Facility, University of Leicester. In brief, GST-tagged apoferritin was overexpressed in *E. coli*, captured on Gluthatione-sepharose beads after cell lysis, cleaved off the resin by TEV protease and purified to homogeneity by size exclusion chromatography in 50 mM Tris-HCl pH 7.5, 100 mM NaCl and 0.5 mM TCEP.

3 µl of apoferritin at 3.8 mg/ml were applied to freshly glow discharged R 1.2/1.3 holey carbon grids (Quantifoil) at 4°C and 100% relative humidity followed by plunge-freezing in liquid ethane using a Vitrobot Mark IV (Thermo Fisher Scientific). The sample concentration resulted in a dense, single-layered hole coverage. Data were collected on a Titan Krios TEM (Thermo Fisher Scientific) operated at 300 kV and a magnification resulting in a calibrated pixel size of 0.834 Å. The energy filter (Gatan) was operated in zero loss mode

with a slit width of 20 eV. The K3 direct electron detector (Gatan) was operated in counting mode with a freshly acquired reference for gain correction. The exposure rate was adjusted to 20 e⁻/px/s. SerialEM⁶⁹ was used for frame and tilt series acquisition.

Positions for both data sets were selected to be distributed evenly over the same grid area to maximize the similarity in ice thickness and particle density. For AF-f, 150 frame series were collected with a total series exposure of 32 e⁻/Å², fractionated in 40 frames. For AF-t, 135 tilt series ranging from -40 to +40 degrees were collected in a grouped dose-symmetric scheme¹⁰⁶ with a group size of 2 and in 2 degree steps. Each tilt was exposed to 2.7 e⁻/Å², fractionated in 3 frames.

3.2.12 Comparison between frame and tilt series performance

Using data set AF-f, frame series alignment and local CTF estimation were performed in Warp with a spatial resolution of 8x5, owing to the rectangular format of the K3 chip. 22,122 particles were picked with a retrained BoxNet model in Warp and exported at full resolution in 512 px boxes. Global 3D refinement with octahedral symmetry was performed in RELION 3.0. The results were imported in *M*. The particle poses were given a temporal resolution of 3, the deformation grid resolution was set to 6x4, and refinement of all parameters was performed for 5 iterations (**Table S3.1**). Data-driven weight estimation was performed to assign unique weights to every series and frame index.

Using data set AF-t, tilt movie frame alignment was performed in Warp using a model without spatial resolution. Initial tilt series alignment was performed in IMOD using patch tracking on 6x binned images with default settings. Tilt series CTF estimation was performed in Warp. 18,991 particles were picked using Warp's 3D template matching in full tomograms reconstructed at 10 Å/px. Sub-tomograms and 3D CTF volumes were exported at 2 Å/px using 140 px boxes. Global 3D refinement with octahedral symmetry was performed in RELION 3.0. The results were imported in *M*. The particle poses were given a temporal resolution of 3, the image warp grid resolution was set to 6x4x41, and refinement of all parameters was performed for 5 iterations, including tilt movie frame

alignment in the last 2 iterations (**Table S3.1**). Data-driven weight estimation was performed to assign unique weights to every series and tilt index.

3.2.13 Assessment of multi-species refinement

Particles from each frame series of the AF-f data set were split in 5% and 95% sub-populations, resulting in species with 3,710 and 70,497 particles, respectively. Frame alignments and particle poses previously obtained from Warp and RELION were reused. In the first scenario, the 5% species was refined alone. In the second scenario, the 5% species was co-refined with the 95% species. Both species were assumed to be structurally independent and did not contribute particles to each other's reconstructions. For both tested scenarios, a 6x4 starting grid for the deformation was used, the resolution of all species was set to 4.0 Å and only one refinement iteration was performed in *M* to avoid possible benefits from the higher resolution the 95% species would reach after the first iteration.

3.2.14 Comparison with RELION on atomic-resolution frame series data

Frame series data were downloaded for the EMPIAR-10248 entry and pre-processed in Warp. 109,437 particles were exported at 0.6 Å/px using 466 px boxes and refined in RELION. The resulting particle poses and half-maps were imported in *M* and refined for 5 iterations starting with a resolution of 3.0 Å in the first iteration. A starting grid of 4x4 was used for the deformation model, and the number of frames was truncated to 25. All CTF-related parameters were refined, including doming, per-series beam tilt and a 3x3 grid model for local astigmatism (**Table S3.1**). For the last 2 iterations, anisotropic per-series, per-frame B-factor weights were estimated. The final iteration was completed in ca. 24 hours, using 4 GeForce 2080 Ti GPUs. The original mask deposited with EMD-9865 was used to estimate the final resolution.

To analyze the doming behavior, fitted doming model parameters were averaged across the data set. Because doming was fitted after per-particle defocus, which was dominated by frames 3–4 due to weighting, the values were normalized by subtracting those of frame 1 from all. As a larger, planar inclination spanning the field of view was observed in the fits in addition to the more local bending of the center relative to the periphery, a

plane was fitted into each frame's values and subtracted from them before quantifying the doming.

3.2.15 Comparison with other tools for tilt series data refinement

Tilt series data were downloaded for the EMPIAR-10064 entry. Initial tilt series alignment was performed in IMOD using manually picked gold fiducials on 4x binned images with default settings. Tilt series CTF estimation was performed in Warp. 3,566 particles were picked using Warp's 3D template matching in full tomograms reconstructed at 10 Å/px. Sub-tomograms and 3D CTF volumes were exported at 5.0 Å/px. Global 3D refinement reached a resolution of 13 Å. The results were imported in M. The particle poses were given a temporal resolution of 3, the image warp and volume warp grid resolutions were set to 8x8x41 and 4x4x2x20, respectively, and refinement of all parameters was performed for 5 iterations (**Table S3.1**). Data-driven anisotropic weight estimation was performed to assign unique weights to every series and tilt index.

The processing of EMPIAR-10045 tilt series was performed in exactly the same way as described in the previous paragraph for EMPIAR-10064, using 3,058 particles (**Table S3.1**).

Tilt series movie data were downloaded for the EMPIAR-10164 entry. Tilt movie frame alignment was performed in Warp using a model without spatial resolution. Initial tilt series alignment was performed in IMOD using gold fiducials automatically picked in Warp, on 6x binned images with default settings. Tilt series CTF estimation was performed in Warp. 130,658 particles were picked using Warp's 3D template matching with a template derived from EMD-3782 in full tomograms reconstructed at 10 Å/px. Sub-tomograms and 3D CTF volumes were exported at 5 Å/px using 56 px boxes. Global 3D refinement with C6 symmetry was performed in RELION 3.0, and reached the 10 Å Nyquist limit. The results were imported in M. The particle poses were given a temporal resolution of 3, the image warp and volume warp grid resolutions were set to 8x8x41 and 3x3x3x20, and refinement of all parameters was performed for 5 iterations, including tilt movie frame alignment in the last 2 iterations (**Table S3.1**). Data-driven anisotropic

weight estimation was performed to assign unique weights to every series, tilt index and tilt frame index.

3.2.16 Acquisition and refinement of *M. pneumoniae in situ* tilt series data

Data previously used in another study¹⁰⁰ were re-analyzed with the release version of *M. As described there, Mycoplasma pneumoniae* strain M129 (ATCC 29342) cells were grown on 200 mesh gold grids coated with a holey carbon support (R 2/1, Quantifoil). Cells were cultivated at 37 °C in modified Hayflick medium: 14.7 g/L Difco PPLO (Becton Dickinson, USA), 20% (v/v) Gibco horse serum (New Zealand origin, Life Technologies, USA), 100 mM HEPES-Na (pH 7.4), 1% (w/w) glucose, 0.002% (w/w) phenol red and 1,000 U/mL freshly dissolved penicillin G. Chloramphenicol (Cm; Sigma-Aldrich, USA) was added 15 minutes prior to vitrification, at a final concentration of 0.5 mg/ml. Grids were quickly washed with PBS buffer containing 10 nm protein A-conjugated gold beads (Aurion, Netherlands), blotted from the back side for 2 seconds, and plunged into mixed liquid ethane/propane at liquid N₂ temperature with a manual plunger (Max Planck Institute of Biochemistry, Germany). The cryo-EM grids were stored in a sealed box in liquid N₂ before usage.

Tilt series data were collected on a Titan Krios TEM operated at 300 kV (Thermo Fisher Scientific) equipped with a field-emission gun, a Gatan K2 Summit direct detector and a Quantum post-column energy filter (Gatan). Images were recorded in exposure-fractionation, counting mode using SerialEM 3.7.2. Tilt-series were acquired with a dose-symmetric scheme using dedicated scripts¹⁰⁶ with the following settings: TEM in nano-probe mode, magnification 81,000 with a calibrated pixel size of 1.7 Å, energy filter in zero loss mode, defocus range 1.5 to 3.5 µm, tilt range -60° to 60° with 3° tilt increment and constant exposure per tilt, total exposure of 120 e⁻/Å². In total, 65 tilt series were collected from Cm-treated cells.

Raw tilt movies were processed in Warp. *De novo* tilt series alignment was performed in IMOD using gold fiducials picked automatically with Warp's BoxNet, and the results were

imported in Warp, where the tilt series CTFs were estimated. Using full tomograms reconstructed at 10 Å/px, two tomograms were denoised using Warp's Noise2Map tool to pick the ribosome particles manually. Using these coordinates, sub-tomograms were exported from Warp to RELION to obtain an initial reference. This reference was used to perform template matching in Warp at 10 Å/px. In addition, a binary classifier based on a 3D CNN was trained on the 2 manually picked tomograms to remove false positives (membranes, carbon hole edges etc.) from the template matching results. 24,202 particles were obtained this way. Sub-tomograms for all particles were exported from Warp to RELION and aligned against the previously refined low-resolution reference. No classification was performed. The results were imported in *M*. There, global movement and rotation, a 5x5x41 image-space warping grid, a 8x8x2x10 volume-space warping grid, as well as particle pose trajectories with 3 temporal sampling points were refined over 5 iterations (**Table S3.1**). Starting with iteration 3, CTF parameters were also refined. At the beginning of iteration 4, reference-based tilt movie alignment was performed, resulting in a 3.7 Å map. Using the improved alignments, sub-tomograms were reconstructed at 3 Å/px. Classification into 5 classes was performed in RELION. 17,890 particles from the 2 best classes were imported in *M* and refined for another iteration using the same settings to obtain a 3.5 Å map. The final iteration was completed in ca. 6 hours, using 4 GeForce 2080 Ti GPUs. Afterwards, focused refinements were performed in *M* using masks limited to the 30S and 50S subunits, optimizing only image warping and particle poses.

To calculate the Rosenthal–Henderson¹⁰¹ plot, deformation, weighting and CTF parameters from the last iteration of 70S refinement were kept. The number of particles was reduced by excluding entire tilt series from the data set, thus keeping the average particle density per series constant. Resolution was reset to 10 Å at the beginning of each subset's refinement, and only the particle pose trajectories were optimized for 3 iterations.

	Series type	Tilt range	Particles picked	Particles classified	Image warp	Volume warp	Tilt movie alignment	Per-particle trajectory samples	CTF refinement	Weight fitting	Higher-order aberrations	Donning	Iterations	Symmetry	Resolution
EMPIAR-10288	F		1,033,994	149,328	2×2×40			2	✓	✓			5	C1	2.9 Å
EMPIAR-10453	T	±60	38,742	22,998	3×2×41	—		1	✓	✓			4	C3	3.8 Å
AF-t	T	±40	18,991		6×4×41	—	✓	3	✓	✓	✓	✓	5	O	2.3 Å
AF-f	F		22,122		6×4×40			3	✓	✓	✓	✓	5	O	2.3 Å
EMPIAR-10248	F		109,437		4×4×25			2	✓	✓	✓	✓	5	O	1.34 Å
EMPIAR-10064	T	±40	3,566		8×8×41	4×4×2×20		3	✓	✓			5	C1	5.7 Å
EMPIAR-10045	T	±60	3,058		8×8×41	4×4×2×20		3	✓	✓			5	C1	6.0 Å
EMPIAR-10164	T	±60	130,658		8×8×41	3×3×20	✓	3	✓	✓			5	C6	3.0 Å
<i>M. pneumoniae</i>	T	±60	24,202	17,890	5×5×41	8×8×2×10	✓	3	✓	✓			6	C1	3.5 Å

Table S3.1 | Refinement parameters for all data sets.

4. Conclusions and outlook

In this thesis, two new computational tools for cryo-EM were described and evaluated: Warp and M. Together, the tools cover a significant portion of the cryo-EM structure determination pipeline and set new resolution records for tomographic data of *in vitro* and *in situ* samples.

Warp bridges the gap between data acquisition and SPA, providing new algorithms for motion correction, CTF estimation, particle picking, and tomogram reconstruction. Because the new algorithms are more robust, Warp is able to execute them automatically and in parallel with data collection to provide real-time monitoring of sample and data quality. Advances in machine learning largely solve the long-standing problem of particle picking, reaching human-like accuracy even on challenging samples. Transparent handling of defocus gradients in tilted samples allows to use this data collection strategy routinely in case of strongly preferred particle orientation. Frame series and tilt series data can be pre-processed using the same user-friendly interface, making Warp a versatile tool for cryo-EM facilities. Extensive testing on published data sets showed that Warp can achieve equal or significantly better resolution than previous pre-processing pipelines.

M takes over from SPA tools like RELION once particle classes and global alignments have been established. It improves the solution further by going back to the raw frame series or tilt series data that had been previously refined with reference-free algorithms, and uses a new multi-particle framework to perform simultaneous reference-based optimization of all aspects of the sample and imaging models. Sample motion within the field of view is constrained in physically plausible ways, and higher-order optical aberrations can be fitted to achieve atomic resolution on favorable samples. We showed that M can achieve equally high resolution for frame and tilt series, closing the gap that had made tilt series unattractive for high-resolution work. Most importantly, our work showed that unprecedentedly high resolution can be obtained for structures imaged inside cells, opening *in situ* structural biology to new applications such as *de novo* structure determination and structure-based drug design.

4.1 Further development of *Warp* and *M*

Warp and *M* are mature, well-integrated with each other and external programs, and have already impacted many projects in structural biology. They will also provide a solid foundation for future methods that go beyond today's SPA.

Warp's pre-processing automation needs to be integrated with acquisition software like *SerialEM*⁶⁹. Implementing this will make acquisition faster and more efficient, especially for tilt series. Unlike conventional micrographs, tilt series are much slower to acquire, making the loss of tracking or the targeting of an undesirable grid location especially costly. Problems can be avoided if they are noticed during pre-processing and communicated to the acquisition software. Furthermore, on-the-fly tilt series alignment must be accelerated to a level where *Warp* can communicate necessary adjustments to *SerialEM* during stage tilting to avoid additional tracking steps.

Interactive and automated tomogram segmentation must be added to *Warp*. While *in vitro* micrographs rarely require segmentation beyond particle picking, *in situ* tomograms automatically reconstructed in *Warp* hold a great wealth of context that must be analyzed and made available to the user and downstream processing algorithms. Such analysis will include the segmentation of organelles, membranes, extracellular vesicles, and other recurring objects. Having such context will greatly benefit any further analysis of *in situ* data.

Many features remain to be added to *M*. The added ability to refine molecules with helical symmetry will open an important class of structures to high-resolution *in situ* refinement. Furthermore, a neural net-based image similarity metric must be developed that is more robust to non-Gaussian noise than conventional correlation. Integrated in *M*, this metric can enable the refinement of smaller molecules both *in vitro* and *in situ* through improved alignment.

4.2 Central repository for sharing *in situ* cryo-ET data

Our processing of *in situ* data with *Warp* and *M* demonstrated that residue-level resolution can be achieved inside a thick, crowded cell. Ribosomes are special in their molecular mass, rigidity, and abundance in cells, making them an ideal target for *in situ* SPA. Most molecules with a molecular mass over 100 kDa are far less abundant. A single tomogram's field of view will contain only a few particles of a single molecule species on the average, and even less than one copy in many cases.

Despite improvements in resolution, our research showed that thousands of particles are still required to achieve high resolution. Even without factoring in the increased compositional and conformational heterogeneity expected for most proteins *in situ*, this translates to a larger amount of data than any researcher or even a lab can acquire and process within a reasonable amount of time. We also demonstrated that our multi-particle refinement framework benefits from a higher number of particles per field of view. Thus, a rare molecule will achieve higher resolution with the same number of particles when co-refined with other molecule species in the same data using *M*. While thousands of molecules beyond the molecule of interest are present in any researcher's *in situ* data, they are not analyzed and cannot contribute to *M*'s multi-particle refinement, even though they could be of high interest to others. Thus, by pooling their data and analyzing them together, everyone stands to gain not only more particles for their projects, but also improved multi-particle alignments through all the other analyzed molecule species.

Scientists are still very reluctant to share *in situ* data because they can power many more projects beyond the initial study. However, once the few ribosome-like molecules have been researched exhaustively, the two strong incentives outlined above will make the field much more collaborative. This is already common in other "big data" fields such as genomics or proteomics, where sharing large amounts of raw data is common and pooled data sets enable studies of unprecedented scale.

Such collaborations will require powerful data sharing mechanisms and advanced processing workflows. Like the collaborative editing of text or program code, collaborative *in situ* cryo-ET necessitates versioning to keep track of the exact processing of every piece of data. The merging of independent processing branches generates conflicts, e.g. when multiple molecules occupy the same space, which must be resolved with intuitive tools. A foundation for this has already been laid in *M*, which keeps track of the exact changes occurring between refinement iterations and links them cryptographically into a version graph. However, branch merging and conflict resolution are still unsolved problems that will require a significant engineering effort.

At the same time, a central repository must be built to host the pooled data sets and provide an open API for local clients like *M* to retrieve data, and submit updated processing results such as the positions and identities of new molecules, improved alignments, or higher-resolution maps. Because of the integrated annotation, SQL-like search queries can enable scientists outside of structural biology to mine the knowledge for their studies.

New data will be automatically segmented and mined for already known structures using continuously improving deep learning models. Because the addition of new data also affects what can be done with previously refined particles, e.g. the discovery of additional states due to more data, such decisions must be automated. Centralized, continuous processing will likely change how we publish results. Whereas currently maps deposited in the Electron Microscopy Data Bank (EMDB) remain frozen after publication, results will be more fluid in the future, continuing to evolve long after the first publication and enabling valuable meta-studies.

4.3 Resolving compositional and conformational heterogeneity with machine learning

Understanding how biological molecules change over time is key to understanding biology. Cryo-EM is in a unique position to image their full dynamics both *in vitro* and *in situ*.

Unfortunately, current methods are still rooted in crystallographic assumptions about the sample's rigidity, and attempt to sort all changes into discrete classes. Such assumptions are a poor fit for the continuous movements and complex interplay of binding partners. Ever bigger data sets are required to resolve the dynamics with finer granularity, but still yield an incomplete picture in the end.

Computational workarounds such as focused refinement restrict the analysis to a sub-region of the molecule. This works well if a flexibly attached sub-region is independent in its behavior, but fails to capture any co-variance between different regions. Methods like multi-body refinement¹⁰⁷ automate the refinement of several regions based on a crude partitioning provided by the user, but only work for certain tree-like region topologies.

Higher-dimensional embeddings of cryo-EM data with 3D principal component analysis (PCA) or variational autoencoders (VAE) are completely agnostic to the type of heterogeneity they model because they operate on voxel intensities and can capture continuous movement, dissociation, and occupancy changes equally well. The resulting models enable visual, interactive exploration of the full heterogeneity space. However, this does not lead to increased resolution because the algorithm does not know that the different states belong to one molecule.

An algorithm must be developed to combine the advantages of both methods. In a first step, a VAE-based embedding will provide a low-resolution, but high-accuracy model of the heterogeneity space. Then a pseudo-atomic representation (i.e. "connected, moveable voxels") will be fitted to the VAE outputs to encode the same space as explicit spatial movement and occupancy changes. Finally, experimental data can be iteratively fitted with the pseudo-atomic representation, and back-projected into a single, flexible basis to obtain both more precise embeddings and a higher-resolution map.

4.4 Applying machine learning to *in situ* cryo-ET data

In situ cryo-ET data present the most exciting opportunity, but also the biggest technical challenge in structural biology today. Applied to chaotic excerpts of cells filled with

dynamic, interacting molecules, shortcomings of modern SPA methods become most evident. While a sizeable portion of *in vitro* samples could be investigated at high resolution even if methods development stopped today, moving beyond ribosomes and a few similarly favorable targets will be very hard *in situ*. A community-driven data repository can solve the problem of data availability. However, to make sense of the data, the field needs to move beyond classic computer vision and SPA approaches.

In vitro data can be easily tailored to a methods developer's needs, and reliably evaluated by eye. For instance, to come up with ground truth data for a particle picking model, the particles can be picked manually. This is not possible *in situ*, where most particles are very hard to recognize by eye, and to pick exhaustively due to their number. The best way to come up with the ground truth is through simulation. Unfortunately, previous attempts at cryo-EM image simulation have failed to deliver a realistic noise model. While our understanding of the physical imaging model is likely sufficient, it is the lack of imperfections in the simulated sample that makes the result unrealistic. A solution to this might be to simulate *in situ* samples with coarse-grained molecular dynamics and to explore a generative-adversarial network-based approach¹⁰⁸ to image simulation. Therein, two antagonistic networks train each other: a generative model attempts to generate a realistic image, while a discriminative network tries to distinguish between real and generated examples. This will allow the generative model to incorporate all the fine statistical intricacies lacking in classic multi-slice methods, and will facilitate the training of a wide range of algorithms on simulated data.

CNNs solved the problem of reliably picking particles *in vitro*⁸⁴. Through training on large amounts of manually labeled data, these models learned to classify image regions as "particle" and "not particle" with superhuman accuracy. Because *in vitro* samples usually contain only the molecule of interest, this lack of selectivity is desirable as it simplifies training and avoids the discrimination of rare particle orientations. However, if a similar training corpus of *in situ* data could be created, applying the trained model to a cellular tomogram would yield tens of thousands of particles, as the model picks every protein.

Manually labeling data for every new molecule of interest and retraining the model is also highly impractical. Instead, a system must be created that can leverage prior knowledge of a high-resolution structure, and possibly orthogonal data sources such as light microscopy, to guide the segmentation process. This will merge the selectivity of template matching with the robustness and high accuracy of a CNN model.

Using improved methods for particle localization and refinement of heterogeneous structures, researchers will be able to model larger portions of *in situ* data. Growing amounts of shared data and analyzed structures will elevate our understanding of many biological processes to a new level. However, cell biology is too complex to be understood particle-by-particle. In the early days of other fields such as natural language processing, chess programs, or expert systems, large databases of manually curated building blocks were leveraged by algorithms following carefully formulated sets of rules to produce poor results. Cryo-EM map and atomic model databases play a similar role today. The success of massive, differentiable models like GPT-3¹⁰⁹ or AlphaZero¹¹⁰ showed that such systems can incorporate knowledge in a humanly incomprehensible form that is more useful to their function (i.e. weights of a neural net), and recombine it in extremely complex ways to achieve superiority in very difficult domains. Our early attempts at denoising tomograms using deep CNNs indicate that the models learn not only a noise model for the signal, but also recurring structural motives in the data such as membranes or highly abundant proteins, and renders them with increased resolution. The future of *in situ* structural biology likely belongs to a system that models the cell in a “particle-less” way, rather “denoising” every observation to high resolution. To do so efficiently, the model will need to learn the building rules from a vast corpus of unlabeled cryo-ET data: starting with amino acids and secondary structure, individual protein structures, all the way up to how proteins are likely to arrange in cells. Orthogonal omics-scale data sources such as proteomics will help refine these rules. If implemented successfully, such a system would fundamentally transform structural biology.

List of abbreviations

2D	2-dimensional
3D	3-dimensional
Å	Angstrom
API	Application programming interface
BIM	Beam-induced motion
Cm	Chloramphenicol
CNN	Convolutional neural network
Cryo-EM	Cryo-electron microscopy
Cryo-ET	Cryo-electron tomography
CTF	Contrast transfer function
DED	Direct electron detector
DQE	Detective quantum efficiency
e ⁻	Electron (physical particle)
FFT	Fast Fourier transform
FRC	Fourier ring correlation
FSC	Fourier shell correlation
FT	Fourier transform
GUI	Graphical user interface
IFT	Inverse Fourier transform
L-BFGS	Limited-memory Broyden-Fletcher-Goldfarb-Shanno
<i>M. pneumoniae</i>	<i>Mycoplasma pneumoniae</i>
NC	Normalized correlation
NCC	Normalized cross-correlation
PCA	Principal component analysis
PS	Power spectrum
ResNet	Residual network
SGD	Stochastic gradient descent
SNR	Signal-to-noise ratio
SSNR	Spectral signal-to-noise ratio
SPA	Single-particle analysis
STA	Sub-tomogram averaging
TEM	Transmission electron microscope
UI	User interface
VAE	Variational autoencoder
WPF	Windows Presentation Foundation

List of figures and tables

Figure 2.1 Warp handles all pre-processing steps to close a gap in the 2D cryo-EM pipeline.	12
Figure 2.2 User interface of <i>Warp</i>	14
Figure 2.3 Deconvolution and denoising of a low-defocus micrograph.....	15
Figure 2.4 Motion and CTF model fitting by Warp.	16
Figure 2.5 CTF fitting of flat, tilted and tilt series data.	17
Figure 2.6 Automated particle picking with Warp's deep learning-based BoxNet.	18
Figure 2.7 Unbiased particle picking with Warp's BoxNet.	19
Figure 2.8 <i>Warp's</i> 2D pipeline improves cryo-EM density for influenza hemagglutinin.	24
Figure 2.9 Warp's 2D pipeline in combination with RELION 3.0 improves cryo-EM density for β -galactosidase.	26
Figure 2.10 Effect of using the full local 3D CTF for template matching in tomograms.	27
Figure 2.11 Sub-tomogram averaging results obtained by using <i>Warp's</i> tilt series CTF estimation and sub-tomogram export.	28
Figure 2.12 Neural network architecture of BoxNet.	38
Figure 2.13 Examples of data used to train BoxNet.	39
Table 2.1 Experimental and synthetic data used to train the general BoxNet model. .	41
Figure 3.1 The Warp-RELION- <i>M</i> pipeline for frame and tilt series cryo-EM data refinement.	49
Figure 3.2 Multi-particle system modeling and optimization.	51
Figure 3.3 Example of a parameter grid pyramid that models in-plane motion in a frame series.	52
Figure 3.4 Benefits of considering more particles per micrograph through multi-species refinement.	53
Figure 3.5 CTF correction at low and high defocus.	55

Figure 3.6 Examples of anisotropic B-factor weighting	57
Figure 3.7 Effects of deep learning-based denoising of reconstructions during refinement.....	59
Figure 3.8 Contributions of individual multi-particle system model components to map resolution.....	60
Figure 3.9 <i>M</i> achieves similar resolution for frame series and tilt series data.....	61
Figure 3.10 Comparison with RELION on atomic-resolution frame series data.....	62
Figure 3.11 Quantification of the doming effect in frame series data.	63
Figure 3.12 Comparison of maps obtained from published tilt series using <i>M</i> or other software.....	64
Figure 3.13 Overview of the <i>M. pneumoniae</i> data set.	66
Figure 3.14 <i>M. pneumoniae</i> 70S ribosome-antibiotic map at 3.5 Å refined with the new Warp–RELION– <i>M</i> pipeline.	68
Table S3.1 Refinement parameters for all data sets.	91

References

1. Dubochet, J., Lepault, J., Freeman, R., Berriman, J.A. & Homo, J.C. Electron microscopy of frozen water and aqueous solutions. *J Microsc* **128**, 219–237 (1982).
2. Nakane, T. et al. Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
3. Glaeser, M., Robert & Hall, J., Richard Reaching the Information Limit in Cryo-EM of Biological Macromolecules: Experimental Aspects. *Biophysical Journal* **100**, 2331-2337 (2011).
4. Scheres, S.H. in *J Mol Biol*, Vol. 415 406-418 (2012).
5. Frank, J. Averaging of low exposure electron micrographs of non-periodic objects. *Ultramicroscopy* **1**, 159-162 (1975).
6. Glaeser, R.M. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J Ultrastruct Res* **36**, 466-482 (1971).
7. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
8. Ruskin, R.S., Yu, Z. & Grigorieff, N. Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of structural biology* **184**, 385-393 (2013).
9. Brilot, A.F. et al. Beam-induced motion of vitrified specimen on holey carbon film. *Journal of structural biology* **177**, 630-637 (2012).
10. Campbell, M.G. et al. Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure (London, England : 1993)* **20**, 1823-1828 (2012).
11. Frank, J. The Envelope of Electron Microscopic Transfer Functions for Partially Coherent Illumination. *Optik* **38**, 519–536 (1973).
12. Danev, R., Tegunov, D. & Baumeister, W. in *eLife*, Vol. 6 (2017).
13. Downing, K. & Glaeser, R. Restoration of weak phase-contrast images recorded with a high degree of defocus: the "twin image" problem associated with CTF correction. *Ultramicroscopy* **108**, 921–928 (2008).
14. Zivanov, J., Nakane, T. & Scheres, S.H.W. Estimation of High-Order Aberrations and Anisotropic Magnification From cryo-EM Data Sets in RELION-3.1. *IUCrJ* **7**, 253-267 (2020).
15. Noble, A.J. et al. Routine Single Particle CryoEM Sample and Grid Characterization by Tomography. *bioRxiv* (2017).
16. Mahamid, J. et al. Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science (New York, N.Y.)* **351**, 969-972 (2016).
17. Lyumkis, D., Brilot, A.F., Theobald, D.L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *Journal of structural biology* **183**, 377-388 (2013).

18. Knauer, V., Hegerl, R. & Hoppe, W. Three-dimensional reconstruction and averaging of 30 S ribosomal subunits of *Escherichia coli* from electron micrographs. *Journal of molecular biology* **163**, 409-430 (1983).
19. Danev, R., Yanagisawa, H. & Kikkawa, M. Cryo-Electron Microscopy Methodology: Current Aspects and Future Directions. *Trends in biochemical sciences* **44**, 837-848 (2019).
20. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J Biol Chem* **294**, 5181-5197 (2019).
21. Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* **10**, 584-590 (2013).
22. Grant, T. & Grigorieff, N. Automatic estimation and correction of anisotropic magnification distortion in electron microscopes. *Journal of structural biology* **192**, 204-208 (2015).
23. Zheng, S.Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332 (2017).
24. Rubinstein, J.L. & Brubaker, M.A. Alignment of cryo-EM movies of individual particles by optimization of image translations. *Journal of structural biology* **192** (2015).
25. McLeod, R.A., Kowal, J., Ringler, P. & Stahlberg, H. Robust image alignment for cryogenic transmission electron microscopy. *Journal of structural biology* **197**, 279-293 (2017).
26. Fernandez, J.-J., Li, S. & Agard, D.A. Consideration of sample motion in cryo-tomography based on alignment residual interpolation. *Journal of structural biology* **205**, 1-6 (2019).
27. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of structural biology* **192**, 216-221 (2015).
28. Bell, J.M., Chen, M., Baldwin, P.R. & Ludtke, S.J. High resolution single particle refinement in EMAN2.1. *Methods (San Diego, Calif.)* **100**, 25-34 (2016).
29. Zhang, K. in *J Struct Biol*, Vol. 193 1-12 (2016).
30. Tan, Y.Z. et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Methods* **14**, 793-796 (2017).
31. Frank, J. Single-Particle Reconstruction of Biological Molecules—Story in a Sample (Nobel Lecture). *Angewandte Chemie International Edition* **57**, 10826-10841 (2018).
32. Voss, N., Yoshioka, C., Radermacher, M., Potter, C. & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *Journal of structural biology* **166**, 205-213 (2009).
33. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
34. Scheres, S.H. Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of structural biology* **189**, 114-122 (2015).

35. Roseman, A.M. FindEM--a fast, efficient program for automatic selection of particles from electron micrographs. *Journal of structural biology* **145**, 91-99 (2004).
36. Chen, J.Z. & Grigorieff, N. SIGNATURE: a single-particle selection system for molecular electron microscopy. *Journal of structural biology* **157**, 168-173 (2007).
37. Scheres, S.H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology* **180**, 519-530 (2012).
38. Bharat, T.A., Russo, C.J., Lowe, J., Passmore, L.A. & Scheres, S.H. Advances in Single-Particle Electron Cryomicroscopy Structure Determination applied to Subtomogram Averaging. *Structure (London, England : 1993)* **23**, 1743-1753 (2015).
39. van Heel, M.K., W.; Schutter, W.; van Bruggen E.F.J. Arthropod hemocyanin studies by image analysis. *Structure and Function of Invertebrate Respiratory Proteins, EMBO Workshop 1982, E.J. Wood* **Life Chemistry Reports**, 69–73 (1982).
40. Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290 (2017).
41. Galaz-Montoya, J.G. & Ludtke, S.J. The advent of structural biology in situ by single particle cryo-electron tomography. *Biophysics Reports* **3**, 17-35 (2017).
42. Zivanov, J., Nakane, T. & Scheres, S.H.W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. *IUCr* **6**, 5-17 (2019).
43. Himes, B.A. & Zhang, P. emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. *Nat Methods* **15**, 955-961 (2018).
44. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* **65**, 386-408 (1958).
45. Kiefer, J. & Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *Ann. Math. Statist.* **23**, 462-466 (1952).
46. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
47. Baydin, A.G., Pearlmutter, B.A., Radul, A.A. & Siskind, J.M. Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.* **18**, 5595–5637 (2017).
48. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. Learning representations by back-propagating errors. *Nature* **323**, 533-536 (1986).
49. Lecun, Y. et al. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1**, 541-551 (1989).
50. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015).
51. He, K., Zhang, X., Ren, S. & Sun, J. in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770-778 (2016).
52. Buchholz, T.-O., Jordan, M., Pigino, G. & Jug, F. in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) 502-506 (Venice, Italy; 2019).
53. Dong, C., Loy, C.C., He, K. & Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**, 295-307 (2016).

54. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278-2324 (1998).
55. Krizhevsky, A., Sutskever, I. & Hinton, G.E. in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 1097-1105 (2012).
56. Wang, F. et al. DeepPicker: A deep learning approach for fully automated particle picking in cryo-EM. *Journal of structural biology* **195**, 325-336 (2016).
57. Abadi, M. et al. in Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation 265-283 (2016).
58. Iudin, A., Korir, P.K., Salavert-Torres, J., Kleywegt, G.J. & Patwardhan, A. in *Nat Methods*, Vol. 13 387-388 (United States; 2016).
59. Berman, H.M. et al. in *Nucleic Acids Res*, Vol. 28 235-242 (2000).
60. Wagner, T. et al. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. *Communications Biology* **2**, 218 (2019).
61. Tagari, M., Newman, R., Chagoyen, M., Carazo, J.M. & Henrick, K. New electron microscopy database and deposition system. *Trends in biochemical sciences* **27**, 589 (2002).
62. Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18037-18041 (2013).
63. Bartesaghi, A. et al. 2.2 Å resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. *Science (New York, N.Y.)* **348**, 1147-1151 (2015).
64. Bharat, T.A. & Scheres, S.H. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nature protocols* **11**, 2054-2065 (2016).
65. Turonova, B., Schur, F.K.M., Wan, W. & Briggs, J.A.G. Efficient 3D-CTF correction for cryo-electron tomography using NovaCTF improves subtomogram averaging resolution to 3.4 Å. *Journal of structural biology* **199**, 187-195 (2017).
66. Nocedal, J. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation* **35**, 773-773 (1980).
67. Sorzano, C.O., Otero, A., Olmos, E.M. & Carazo, J.M. Error analysis in the determination of the electron microscopical contrast transfer function parameters from experimental power Spectra. *BMC structural biology* **9**, 18 (2009).
68. Penczek, P.A. et al. CTER—Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy* **140**, 9-19 (2014).
69. Mastronarde, D.N. Automated electron microscope tomography using robust prediction of specimen movements. *Journal of structural biology* **152**, 36-51 (2005).
70. Voortman, L.M., Stallinga, S., Schoenmakers, R.H.M., Vliet, L.J.v. & Rieger, B. A fast algorithm for computing and correcting the CTF for tilted, thick specimens in TEM. *Ultramicroscopy* **111**, 1029-1036 (2011).

71. Schur, F.K. et al. An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. *Science (New York, N.Y.)* **353**, 506-508 (2016).
72. Xiong, Q., Morpew, M.K., Schwartz, C.L., Hoenger, A.H. & Mastronarde, D.N. CTF determination and correction for low dose tomographic tilt series. *Journal of structural biology* **168**, 378-387 (2009).
73. Hutchings, J., Stancheva, V., Miller, E.A. & Zanetti, G. Subtomogram averaging of COPII assemblies reveals how coat organization dictates membrane shape. *Nat Commun* **9** (2018).
74. Russo, C.J. & Henderson, R. Ewald sphere correction using a single side-band image processing algorithm. *Ultramicroscopy* **187**, 26-33 (2018).
75. Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *Journal of structural biology* **157**, 117-125 (2007).
76. Kunz, M. & Frangakis, A.S. Three-dimensional CTF correction improves the resolution of electron tomograms. *Journal of structural biology* **197**, 114-122 (2017).
77. Heymann, J.B., Chagoyen, M. & Belnap, D.M. Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology. *Journal of structural biology* **151**, 196-207 (2005).
78. Vulovic, M. et al. Image formation modeling in cryo-electron microscopy. *Journal of structural biology* **183**, 19-32 (2013).
79. Rickgauer, J.P., Grigorieff, N. & Denk, W. Single-protein detection in crowded molecular environments in cryo-EM images. *eLife* **6** (2017).
80. Mao, X.-J., Shen, C. & Yang, Y.-B. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. *arXiv* (2016).
81. Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)* **36**, 107 (2017).
82. Lehtinen, J. et al. Noise2Noise: Learning Image Restoration without Clean Data. *arXiv*, 1803.04189 (2018).
83. Kremer, J.R., Mastronarde, D.N. & McIntosh, J.R. Computer visualization of three-dimensional image data using IMOD. *Journal of structural biology* **116**, 71-76 (1996).
84. Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with Warp. *Nat Methods* **16**, 1146-1152 (2019).
85. Scheres, S.H. Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol* **579**, 125-157 (2016).
86. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018).
87. Saxton, W.O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J Microsc* **127**, 127-138 (1982).
88. Russo, C. & Henderson, R. Ewald Sphere Correction Using a Single Side-Band Image Processing Algorithm. *Ultramicroscopy* **187**, 26-33 (2018).

89. Cardone, G., Heymann, J.B. & Steven, A.C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *Journal of structural biology* **184**, 226-236 (2013).
90. Scheres, S.H. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat Methods* **9**, 853-854 (2012).
91. Krishna Kumar, K. et al. Structure of a Signaling Cannabinoid Receptor 1-G Protein Complex. *Cell* **176**, 448-458.e412 (2019).
92. Turoňová, B. et al. In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. (2020).
93. Ramlaul, K., Palmer, C.M., Nakane, T. & Aylett, C.H.S. Mitigating Local Over-fitting During Single Particle Reconstruction with SIDESPLITTER. *Journal of structural biology* **211** (2020).
94. Punjani, A., Zhang, H. & Fleet, D.J. Non-uniform refinement: Adaptive regularization improves single particle cryo-EM reconstruction. *bioRxiv*, 2019.2012.2015.877092 (2019).
95. Eisenstein, F., Danev, R. & Pilhofer, M. Improved applicability and robustness of fast cryo-electron tomography data acquisition. *Journal of structural biology* **208**, 107-114 (2019).
96. Kato, T. et al. CryoTEM with a Cold Field Emission Gun That Moves Structural Biology into a New Stage. *Microscopy and Microanalysis* **25**, 998-999 (2019).
97. Chen, M. et al. A complete data processing workflow for cryo-ET and subtomogram averaging. *Nat Methods* **16**, 1161-1168 (2019).
98. Khoshouei, M., Pfeffer, S., Baumeister, W., Forster, F. & Danev, R. Subtomogram analysis using the Volta phase plate. *Journal of structural biology* **197**, 94-101 (2017).
99. Himes, B.A. emClarity Wiki, change log, "2018-Nov-14" entry. <https://github.com/bHimes/emClarity/wiki>. (2020).
100. O'Reilly, F.J. et al. In-cell architecture of an actively transcribing-translating expressome. *Science (New York, N.Y.)* **369**, 554-557 (2020).
101. Rosenthal, P.B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *Journal of molecular biology* **333**, 721-745 (2003).
102. Git – free and open source distributed version control system. <https://git-scm.com>. (2020).
103. Mahamid, J. et al. Visualizing the Molecular Sociology at the HeLa Cell Nuclear Periphery. *Science (New York, N.Y.)* **351**, 969-972 (2016).
104. DeRosier, D. Correction of High-Resolution Data for Curvature of the Ewald Sphere. *Ultramicroscopy* **81**, 83-98 (2000).
105. Schilbach, S. et al. Structures of transcription pre-initiation complex with TFIID and Mediator. *Nature* **551**, 204-209 (2017).
106. Hagen, W., Wan, W. & Briggs, J. Implementation of a Cryo-Electron Tomography Tilt-Scheme Optimized for High Resolution Subtomogram Averaging. *Journal of structural biology* **197** (2017).

107. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S.H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife* **7** (2018).
108. Goodfellow, I.J. et al. in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 2672–2680 (MIT Press, Montreal, Canada; 2014).
109. Brown, T.B. et al. Language Models are Few-Shot Learners. *ArXiv e-prints* (2020).
110. Silver, D. et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv e-prints* (2017).